Análise de Influência na Regressão em Cristas

Koki Fernando Oikawa

DISSERTAÇÃO APRESENTADA AO

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DA UNIVERSIDADE DE SÃO PAULO

PARA OBTENÇÃO DO TÍTULO DE

MESTRE EM ESTATÍSTICA

Área de Concentração: **ESTATÍSTICA**

Orientadora: **Profa. Dra. Silvia Nagib Elian**

- São Paulo, Agosto de 2008 -

Análise de Influência na Regressão em Cristas

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Koki Fernando Oikawa
e aprovada pela Comissão Julgadora.

Banca Examinadora:

- Profa. Dra. Sílvia Nagib Elian (orientadora) – IME-USP.
- Prof. Dr. Wilton de Oliveira Bussab – Fundação Getúlio Vargas.
- Prof. Dr. Luiz Koodi Hotta – IMECC – UNICAMP.

Agradecimentos

Não posso me esquecer, aqui, de agradecer duas grandes mulheres que, corajosamente, me aceitaram em suas vidas e que, quase certamente, não sabem o quanto me fazem bem: Vera Lúcia e Dona Marcelina. Também à minha mãe, minha avó e ao meu tio Nico.

Agradecimentos especiais à minha orientadora Silvia N. Elian por, entre outras coisas, me ajudar a recuperar um pouco de autoconfiança.

Aos meus colegas da Unicapital: Farina, José Roberto e demais professores e também ao meu colega Sidney Miranda pela ajuda de última hora.

Aos meus amigos com os quais sempre aprendo alguma coisa: Camarço, Fábio, Edney, João Paulo, Parmigiani e Zé Américo.

Jamais poderia finalizar essa seção sem ao menos fazer uma menção honrosa ao Jean Paul Sartre e à Simone de Beauvoir, por colocarem minha cabeça no lugar, e ao cineasta mais genial que já existiu: José Mojica Marins.

Resumo

Modelos de Regressão em Cristas, embora possam ser considerados como casos particulares do modelo de regressão linear geral, apresentam características próprias e problemas específicos. São geralmente utilizados para contornar o problema da multicolinearidade, conseqüência da existência de relações lineares entre as variáveis explicativas. No presente trabalho nos dedicamos, inicialmente, à discussão do problema da multicolinearidade induzida por pontos discrepantes. Serão analisados alguns procedimentos para auxiliar a identificação de multicolinearidade gerada por pontos discrepantes. Medidas de influência adaptadas ao contexto de regressão em cristas serão apresentadas, bem como medidas de influência local. O modelo robusto em cristas será abordado e, finalmente, alguns dos procedimentos descritos serão aplicados em um conjunto de dados reais.

Abstract

Ridge Regression Models, even so can be considered as a particular case of the general linear regression model, they present proper characteristics and specific problems. These models are used, in general, to solve the problem of multicollinearity, which is a consequence of existence of linear relation among regressor variables. In the present work, we dedicate, initially, to the discussion of the problem of outlier-induced multicollinearity. Some procedures will be analyzed as helpful recommendations for outlier-identification techniques. Influence measures in ridge regression and local influence approaches will be presented. Robust ridge regression model will be treated and, finally, some of the described procedures will be applied to a real data set.

Sumário

Capítulo 1

Introdução

Modelos de regressão linear múltipla são frequentemente utilizados na análise da relação de uma variável resposta com um conjunto de variáveis explicativas.

Dentro desse contexto, o método de mínimos quadrados é a forma mais comum de obtenção de estimadores dos parâmetros e, satisfeitas as suposições básicas, o procedimento apresenta boas propriedades.

Mas, frequentemente, o problema da multicolinearidade, ou seja, a existência de relações entre as variáveis explicativas, está presente nos dados reais. Esse fato invariavelmente compromete os resultados obtidos pelo método dos mínimos quadrados pois, por exemplo, eleva consideravelmente a variância dos estimadores dos seus coeficientes de regressão. Outras sérias conseqüências podem, ainda, ser destacadas.

Nesse sentido, os estimadores em cristas (*ridge*) surgiram com o objetivo explícito de contornar o problema da multicolinearidade. Dessa forma, modelos de regressão em cristas seriam ajustados para, posteriormente, serem analisados.

Nesse trabalho, nos dedicaremos ao estudo de medidas de influência para o procedimento de regressão em cristas. Tais medidas teriam a finalidade de identificar observações influentes quando os parâmetros do modelo de regressão linear são estimados através do procedimento de regressão em cristas. Iniciaremos o trabalho com a apresentação de uma situação específica, qual seja, aquela na qual a multicolinearidade é induzida por pontos discrepantes. Isto será feito no Capítulo 2.

No Capítulo 3 apresentaremos uma adaptação das medidas de influência tradicionais ao contexto de regressão em cristas.

O Capítulo 4 discute as chamadas medidas de influência local, que se diferenciam daquelas descritas no capítulo anterior pelo fato de serem aplicáveis apenas em procedimentos de estimação por meio de função de verossimilhança.

No Capítulo 5, o método robusto de estimação em cristas será apresentado como alternativa à busca e retirada de observações discrepantes. Finalmente, no Capítulo 6, serão realizadas algumas aplicações em um conjunto de dados reais.

Capítulo 2

Multicolinearidade Gerada por Pontos Discrepantes

2.1 - Introdução

O objetivo principal desse capítulo será o de analisar um particular tipo de multicolinearidade, que é aquela gerada especificamente por pontos discrepantes e avaliar o efeito que esse tipo de problema pode exercer nos estimadores dos coeficientes do modelo de regressão. Em seguida, serão propostos alguns procedimentos para detectar esses pontos discrepantes.

Apresentaremos, inicialmente, uma breve descrição dos conceitos básicos necessários ao desenvolvimento do capítulo e do restante do trabalho.

2.2 – Multicolinearidade e Estimadores em Cristas

Um modelo de regressão linear múltipla com variável resposta Y e variáveis explicativas $x_1, x_2, \ldots, x_p$, não aleatórias, pode ser descrito pela equação:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon .$$

Em linguagem matricial, esse modelo pode ser reescrito na forma linear geral como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

onde

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

O vetor-coluna $\mathbf{y}$, de ordem $n \times 1$, corresponderia ao conjunto de n observações da variável resposta. A matriz $\mathbf{X}$, de ordem $n \times (p+1)$, representa as observações das variáveis explicativas e a coluna de números 1 corresponde ao intercepto. O vetor coluna $\boldsymbol{\beta}$ é o vetor de parâmetros a serem estimados enquanto que o vetor coluna $\boldsymbol{\varepsilon}$ representa o vetor de erros não observados.

Supomos, ainda, que $\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2 \cdot \mathbf{I}_n\right)$, o que implica $E(\boldsymbol{\varepsilon}) = 0$ e que

$$V(\boldsymbol{\varepsilon}) = \sigma^2 \cdot \mathbf{I}_n = \sigma^2 \cdot \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}.$$

Um modelo de regressão ajustado correspondente às variáveis explicativas pode ser escrito como

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}},$$

onde o vetor de estimadores $\hat{\boldsymbol{\beta}}$ é representado por

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}.$$

A diferença entre cada valor observado y_i e seu respectivo valor ajustados $\hat{y}_i$ é o resíduo $e_i = y_i - \hat{y}_i$, sendo que

$$\mathbf{e} = y - \hat{y} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

representa o vetor de resíduos.

2.2.1 – Estimadores de Mínimos Quadrados em Regressão Linear Múltipla

Constitui um resultado conhecido na literatura que o estimador de mínimos quadrados para $\boldsymbol{\beta}$ é descrito por

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Seu valor esperado é denotado por

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta},$$

enquanto que sua matriz de covariância é

$$V(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Como podemos perceber, o estimador de mínimos quadrados $\hat{\beta}$ depende da matriz inversa de $\mathbf{X'X}$. No caso de haver uma perfeita dependência linear entre as variáveis explicativas, a inversa da matriz $\mathbf{X'X}$ não existirá e o estimador $\hat{\beta}$ não poderá ser obtido.

Quando existe uma forte dependência linear, apesar de não ser uma perfeita dependência linear, ainda assim o estimador $\hat{\beta}$ pode ser obtido mas, como conseqüência direta do problema da multicolinearidade, uma série de problemas surgirá.

Multicolinearidade, basicamente, diz respeito ao problema da dependência entre as variáveis explicativas ou, em outras palavras, reflete a existência de combinação linear entre colunas de X e, portanto, de $X'X$.

São várias as possíveis fontes da multicolinearidade e em situações como essa, a utilização do estimador $\hat{\beta}$ não é aconselhável.

Um dos problemas causados pela multicolinearidade faz com que, por exemplo, pequenas mudanças nos dados amostrais produzam grandes alterações sejam de magnitudes, sejam de sinais, no próprio estimador $\hat{\beta}$ de mínimos quadrados. Em outras palavras, o estimador apresentará um comportamento instável em função da existência de multicolinearidade.

2.2.2 – O Estimador em Cristas

O estimador em cristas (estimador ridge) surgiu como uma forma de contornar o problema de multicolinearidade que pode aparecer em dados amostrais.

O estimador em cristas, originalmente proposto por Hoerl e Kennard (1970), é descrito por

$$\hat{\boldsymbol{\beta}}_{\mathfrak{R}} = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_n)^{-1}\mathbf{X}'\mathbf{y}.$$

A diferença entre o estimador em cristas e o de mínimos quadrados reside na soma de uma constante $k \geq 0$ (geralmente pequena) à diagonal principal da matriz $\mathbf{X}'\mathbf{X}$.

Notamos que no caso em que $k = 0$, o estimador $\hat{\boldsymbol{\beta}}_{\mathfrak{R}}$ retorna à sua expressão conhecida de mínimos quadrados $\hat{\boldsymbol{\beta}}$.

Pode-se dizer que o problema com o estimador de mínimos quadrados, num contexto de multicolinearidade, é que embora seja não viciado, sua variância é grande.

Por outro lado, o estimador em cristas é viciado mas seu erro quadrático médio pode ser menor que o do estimador não viciado $\hat{\boldsymbol{\beta}}$ de mínimos quadrados (Montgomery e Peck, 1982, pg. 311), devido ao decréscimo na variância.

Existem inúmeros critérios para a determinação do valor de k e vários deles estão descritos em Oishi (1983).

2.3 – Diagnóstico em Modelos de Regressão

Em algumas análises, as estatísticas básicas podem mudar muito quando um elemento amostral é retirado. Esse ponto será denominado "influente".

Nesta seção, descreveremos brevemente algumas técnicas de diagnóstico com o objetivo de detectar pontos influentes no modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ com $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$.

Ao estimar $\boldsymbol{\beta}$ através de $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, estimador de mínimos quadrados, o vetor de valores ajustados para $\mathbf{y}$, $\hat{\mathbf{y}}$, é dado por:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{H}\mathbf{y}.\end{aligned}$$

A matriz $\mathbf{H}$, definida por $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, também conhecida como matriz "hat", é simétrica e idempotente $(\mathbf{H}^2 = \mathbf{H})$ e de grande utilidade na análise de diagnóstico.

Os elementos da matriz $\mathbf{H}$, h_{ij}, são dados por

$$h_{ij} = \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j,$$

enquanto que os elementos da diagonal principal h_{ii}'s são dados por

$$h_{ii} = \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i,$$

onde $\mathbf{x}'_i$ e $\mathbf{x}'_j$ são, respectivamente, a *i-ésima* linha e a *j-ésima* linha da matriz $\mathbf{X}$.

Acrescenta-se também que $\sum_{i=1}^{n} h_{ii} = p+1$ e que $0 \le h_{ii} \le 1$. Dessa maneira, o "valor médio de h_{ii}" é igual a $\frac{p+1}{n}$.

Uma possibilidade de identificação de pontos influentes sugere que se dê especial atenção para os pontos $\mathbf{x}_i$ (i-ésima linha de $\mathbf{X}$) tais que $h_{ii} > \frac{2(p+1)}{n}$.

Em particular, para $p > 2$, é difícil visualizar pontos $\mathbf{x}_i$ distantes do grupo e a diagonal de $\mathbf{H}$ é uma importante fonte de informação.

É importante observar que nem sempre um resíduo usual associado a um ponto "discrepante" é alto. Com esse objetivo, e também complementando a análise, são definidos vários outros tipos de resíduos que descreveremos a seguir.

Assim como o erro ε_i, verifica-se que cada resíduo e_i também possui média zero, mas variâncias possivelmente desiguais, pois

$$E(\mathbf{e}) = 0 \quad \text{e}$$
$$V(\mathbf{e}) = \sigma^2(\mathbf{I}_n - \mathbf{H}).$$

Dessa maneira, temos que

$$V(e_i) = \sigma^2(1 - h_{ii}), \quad i = 1,\ldots,n.$$

Com relação a algumas de suas classificações, temos que:

- o resíduo $e_i = y_i - \hat{y}_i$, utilizado com muita freqüência, é conhecido como resíduo usual;
- o resíduo padronizado é representado por $z_i = \frac{e_i}{\hat{\sigma}}$, com $\hat{\sigma}^2$ sendo o quadrado médio do resíduo, $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - p - 1}$;
- o resíduo "internamente studentizado" é dado por $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$;
- o resíduo "externamente studentizado" é definido como $t_i = \frac{y_i - \tilde{y}_i}{\hat{\sigma}_{(i)}\left[I_n + \mathbf{X}'_{(i)}\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}\right]^{1/2}}$, com $\tilde{y}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}}_{(i)}$.

Os estimadores $\hat{\boldsymbol{\beta}}_{(i)}$ e $\hat{\sigma}^2_{(i)}$ são obtidos da forma usual, com base na amostra de $n-1$ observações, excluindo a *i-ésima* observação.

Podemos também escrever $t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}$, lembrando que o caso i não entra no cálculo de $\hat{\sigma}_{(i)}$.

Com o objetivo de avaliar a influência de uma particular observação, são ainda construídas medidas de influência.

A mais conhecida delas é a Distância de Cook (D de Cook), que é dada por:

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)' (\mathbf{X}'\mathbf{X}) \left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)}{(p+1)\hat{\sigma}^2}.$$

Como

$$\mathbf{X}\hat{\boldsymbol{\beta}}_{(i)} = \hat{\mathbf{y}}_{(i)}$$

e

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}},$$

temos alternativamente que

$$D_i = \frac{\left(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}}\right)'\left(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}}\right)}{(p+1)\hat{\sigma}^2}.$$

Com o objetivo de detectar influência, a literatura sugere:

- analisar o caso com o mais alto D_i;
- analisar valores de D_i próximos de 1 e maiores ou iguais a 1.

A Distância de Cook D_i também pode ser obtida sob a forma

$$D_i = \frac{1}{p+1} \cdot r_i^2 \cdot \left(\frac{h_{ii}}{1-h_{ii}}\right).$$

A partir dessa expressão, observamos que pontos com alto valor de h_{ii} ou r_i apresentarão grandes valores de D_i.

Uma medida de influência também muito utilizada é a *DFITTS* que é dada por

$$DFITTS_{(i)} = \frac{\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)' (\mathbf{X}'\mathbf{X}) \left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)}{\hat{\sigma}^2_{(i)}},$$

onde $\hat{\sigma}^2_i$ é o estimador de mínimos quadrados de σ^2 sem a *i-ésima* observação.

2.4 – Multicolinearidade Gerada por Pontos Discrepantes

Nesta seção, descreveremos o estudo realizado por Mason e Gunst (1985), que mostra como a multicolinearidade pode surgir em função da existência de pontos discrepantes, bem como seus efeitos nos estimadores dos parâmetros do modelo de regressão. Em seguida, será apresentado um procedimento desenvolvido pelos autores para diagnosticar multicolinearidade especificamente provocada pela existência desses pontos discrepantes.

Na literatura, as principais fontes de multicolinearidade são geralmente resumidas em quatro tipos: i) multicolinearidade devida à restrições nos modelos, ii) características populacionais que restringem os valores das variáveis, iii) deficiências no processo de amostragem e iv) modelos com mais parâmetros que observações. A contribuição de Mason e Gunst (1985) está em retratar uma fonte de multicolinearidade que não é bem conhecida ou tratada na literatura, a saber, multicolinearidade induzida por pontos discrepantes.

Por *outlier* entendemos qualquer observação $(y, \mathbf{x}') = (y, x_1, ..., x_p)$ que distoe radicalmente em relação às demais.

Um *ponto de alavanca* pode ser encarado como um *outlier* entre as variáveis explicativas. Se pontos de alavancagem não se apresentam em grupos, eles são identificados facilmente quando $h_{ii} > 2(p+1)/n$. Mas se esses pontos de alavancagem apresentam-se em grupos, esse método pode ser ineficiente.

A expressão *observação influente* pode ser utilizada para um *outlier* cuja inclusão no conjunto de dados altera substancialmente as estimativas dos coeficientes de regressão, a previsão ou os procedimentos inferenciais associados.

De acordo com Andrade (2004), observações discrepantes podem, de uma forma geral, ser classificadas como aberrantes, influentes ou de alavanca (alto *leverage*), não necessariamente em uma única categoria. Observações aberrantes são aquelas mal ajustadas, caracterizadas por terem resíduos elevados e afetam principalmente o intercepto do modelo. As observações influentes são aquelas que têm um peso desproporcional nas estimativas dos coeficientes. Já os pontos de alavanca são aqueles que têm uma influência desproporcional no próprio valor ajustado, os quais, em geral, estão posicionados em regiões remotas do subespaço gerado pela(s) coluna(s) da matriz de planejamento $\mathbf{X}$. Pontos de alavanca podem também ser influentes, porém, não é comum pontos de alavanca serem aberrantes.

O modelo descrito por Mason e Gunst (1985) é dado por:

$$\mathbf{y} = \beta_0 \mathbf{1} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

onde $\mathbf{y}$ é um vetor n-dimensional da variável resposta, $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_p)$ é a matriz das variáveis explicativas cujos valores são padronizados, isto é, $\mathbf{Z}'\mathbf{Z}$ é a matriz de correlação, β_0 é uma constante desconhecida, $\mathbf{1}$ é um vetor n-dimensional de valores 1, $\boldsymbol{\beta}$ é um vetor p-dimensional de coeficientes desconhecidos de regressão e ε é um vetor n-dimensional de erros aleatórios não observáveis com $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Com relação ao conceito de multicolinearidade, os autores a definem da seguinte maneira:

Definição: Multicolinearidade é tida como existente entre as colunas de $\mathbf{Z}$ se, para algum $\eta > 0$ (pequeno) específico, existir um vetor de constantes $\mathbf{c}' = (\mathbf{c}_1, ..., \mathbf{c}_p)$ nem todos iguais a zero, tais que

$$\sum_{j=1}^{p} \mathbf{c}_j \mathbf{Z}_j = \delta \qquad com \quad \|\delta\| < \eta \cdot \|\mathbf{c}\|, \qquad onde \; \|\mathbf{a}\| = (\mathbf{a}'\mathbf{a})^{1/2}. \tag{2.2}$$

A fim de demonstrar que a multicolinearidade pode não ter necessariamente como origem pontos discrepantes, mas a existência de pontos discrepantes pode implicar em multicolinearidade, a ilustração que segue foi utilizada pelos autores.

Considere inicialmente o caso em que $p=2$. Sejam $\mathbf{s}_i=(s_{i1},s_{i2})'$, $i=1,...,n$, $\mathbf{u}_1^*=\theta\cdot\mathbf{s}_1 \quad (\theta, s_{11}, s_{12}>0)$ e $\mathbf{u}_i^*=\mathbf{s}_i \ (i\neq 1)$, onde $\mathbf{u}_i^{*'}$ representa a *i-ésima* linha da matriz (não padronizada) das variáveis explicativas, excluído o termo constante, qual seja, $\mathbf{x}_i'=(1,\mathbf{u}_i^{*'})$. De acordo com essa construção, fazendo $\theta\to\infty$, as variáveis explicativas referentes à primeira observação assumirão valores extremos em relação às demais observações. Denominemos as linhas de $\mathbf{Z}$ correspondentes a $\mathbf{u}_i^{*'}$ por $\mathbf{u}_i'$.

Por meio do cálculo da inversa de matrizes particionadas, é possível verificar que

$$\begin{aligned} h_{11} &= n^{-1}+\mathbf{u}_1'(Z'Z)^{-1}\mathbf{u}_1 \\ &= n^{-1}\{1+(n-1)\theta^2 d/(1+\theta^2 d)\}, \end{aligned} \tag{2.3}$$

onde $d>0$ é função apenas de $\mathbf{s}_i$. Assim, quando $\theta\to\infty$, $h_{11}\to 1$ e, para θ suficientemente grande, $h_{11}>2(p+1)/n$. Verifica-se ainda que quando $\theta\to\infty$,

$$\begin{aligned} z_{1j} &\to \{(n-1)/n\}^{1/2}, \qquad j=1,2 \\ z_{ij} &\to \{n(n-1)\}^{1/2}, \qquad i\neq 1, \quad j=1,2. \end{aligned} \tag{2.4}$$

Assim, para θ suficientemente grande, $\mathbf{Z}_1-\mathbf{Z}_2$ pode ser arbitrariamente aproximada de zero, assegurando, assim, a existência de multicolinearidade apresentada em (2.2). Dessa forma, a existência desse ponto discrepante, observação 1, provocou o surgimento do problema da multicolinearidade.

Os autores fizeram uso do mesmo tipo de argumento para estender o resultado para o caso de $p=q$ variáveis explicativas.

O problema torna-se mais claro através da análise da Figura 2.1.

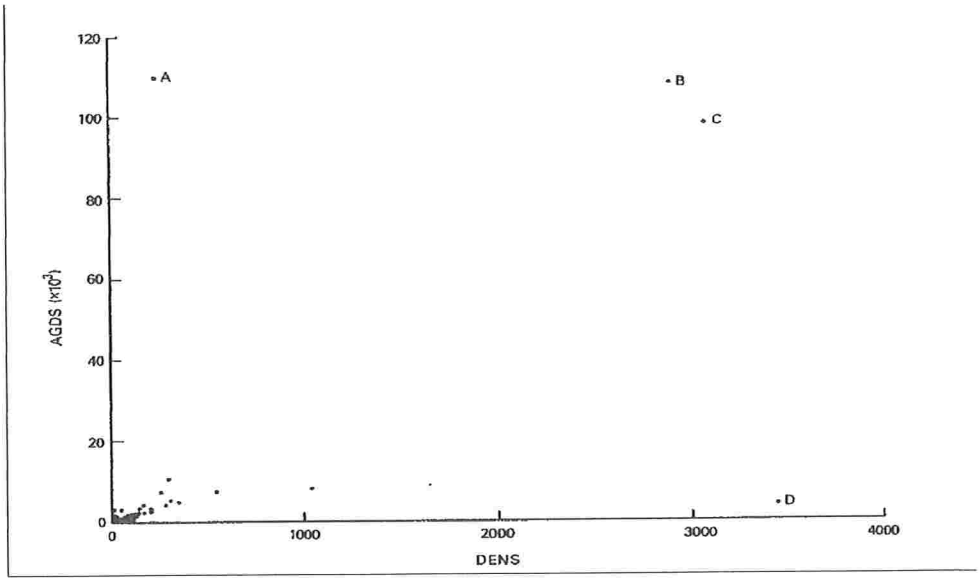


Figura 2.1 - Diagrama de dispersão das variáveis AAC e DENS.
Fonte: Mason e Gunst (1985).

Esta figura refere-se a um trabalho no qual Mason e Gunst (1985) colheram uma amostra de dados do Produto Nacional Bruto (PNB) que, nesse caso, é a variável resposta, para uma amostra de 49 países, com as variáveis sócio-econômicas explicativas a seguir: taxa de mortalidade infantil (TMI), relação médico/população (RMP), densidade populacional (DENS), densidade como função da área agrícola cultivada (AAC), medida de alfabetização (MA) e um índice de educação superior (IES). O objetivo era o de ajustar um modelo de regressão linear da variável logaritmo do PNB em função das variáveis explicativas. Aparentemente, os dados apresentaram multicolinearidade nas variáveis AAC e DENS, fato esse que será descrito futuramente.

A Figura 2.1 apresenta o diagrama de dispersão das variáveis AAC e DENS. Nele, podemos perceber quatro pontos discrepantes: A, B, C e D. Se os dados consistissem de todos os valores exceto os pontos B e C, as observações A e D seriam consideradas apenas como pontos de alavanca que não gerariam multicolinearidade. Se, por outro lado, os dados contivessem todos os pontos exceto A e D, as observações B e

C gerariam multicolinearidade e só seriam detectadas como pontos de alavanca se ocorressem isoladamente. Tal fato é conseqüência da proximidade entre ambas e de sua distância relativa às demais observações. Os pontos B ou C, isoladamente, são exemplos da situação em que as expressões (2.3) e (2.4) atingiram seus valores limite.

Na existência de multicolinearidade gerada por pontos discrepantes, poderíamos inadvertidamente pensar que seus efeitos sobre os estimadores de mínimos quadrados seriam de dois tipos: i) aqueles geralmente associados ao problema da multicolinearidade e ii) aqueles efeitos decorrentes da existência de pontos discrepantes. Verifica-se, no entanto, que esses efeitos podem ser muito diferentes.

De acordo com Hoerl e Kennard (1970), citado em Mason e Gunst (1985), a multicolinearidade tende a produzir valores grandes para as estimativas de mínimos quadrados, enquanto que Dorsett e Gunst (1982) atestam que pontos discrepantes podem fazer com que estimativas de mínimos quadrados tendam para zero.

Assim, diagnosticar a causa da multicolinearidade passa a ser tão importante quanto detectá-la e um exame cuidadoso deve ser feito acerca da natureza da multicolinearidade contida nos dados.

No entanto, conforme discutido no exemplo, a presença de outliers múltiplos pode dificultar a verificação.

Como já comentado anteriormente, estimadores viciados surgem, frequentemente, como alternativas para o problema da multicolinearidade. Mas esta não será uma boa estratégia se a multicolinearidade for gerada por pontos discrepantes.

Para ilustrar essa idéia considere, novamente, o exemplo em que $p = 2$ e $u_1^* = \theta \cdot \mathbf{s}_1$.

Podemos propor, neste caso, o uso do estimador em cristas para $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{\mathfrak{R}} = \left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}\right)^{-1}\mathbf{Z}'\mathbf{y}, \qquad k > 0. \tag{2.5}$$

Aplicando o resultado (2.4), Mason e Gunst (1985) demonstraram que para n grande e quando $\theta \to \infty$,

$$\hat{\boldsymbol{\beta}}_{\mathfrak{R}} \to (k+p)^{-1}(y_1 - \bar{y}_{(1)})\mathbf{s}^*, \qquad para \qquad \bar{y}_{(1)} = (n-1)^{-1}\sum_{i \neq 1} y_i, \tag{2.6}$$

e $\mathbf{s}^*$ um vetor bi-dimensional contendo os sinais dos elementos de $\mathbf{s}_1$. Assim, ambos os elementos do estimador em cristas em (2.6) possuem o mesmo valor e seus sinais relativos são determinados pelos sinais dos elementos em $\mathbf{s}_1$, embora nenhuma restrição nesse sentido tenha sido imposta ou esperada.

Observa-se, portanto, a inadequação do uso deste estimador e, segundo os autores, outros possíveis estimadores viciados sofreriam do mesmo problema.

Voltando à discussão do conjunto de dados da Figura 2.1, com o objetivo de avaliar a multicolinearidade entre as variáveis explicativas, os autores calcularam o fator de inflação da variância (FIV), autovalores e autovetores da matriz de correlação das variáveis explicativas.

Para uma melhor compreensão do FIV, considere R_j^2 o coeficiente de explicação do modelo de regressão linear de $\mathbf{X}_j$ em função de $\mathbf{X}_1, ..., \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, ..., \mathbf{X}_p$. Se a_{jj} é o j-ésimo elemento da matriz $(\mathbf{X'X})$, para as variáveis padronizadas, verifica-se que $a_{jj} = \dfrac{1}{1-R_j^2}$. Assim, se $\mathbf{X}_j$ for altamente correlacionada com as demais variáveis explicativas, $R_j^2 \approx 1$ e $Var(\hat{\boldsymbol{\beta}}_j) = a_{jj}\sigma^2$ será grande. Por esse motivo, a_{jj} é denominado Fator de Inflação da Variância. A literatura sugere que se FIV > 5, a estimativa do correspondente coeficiente da variável está com problemas devido à multicolinearidade. Acrescenta-se também que VIF > 5 implica em $R_j^2 > 0{,}8$.

Na mesma direção, os autovalores $\lambda_1,...,\lambda_p$ da matriz $(\mathbf{X}'\mathbf{X})$ e seus respectivos autovetores também podem ser usados com a finalidade de se detectar multicolinearidade. Se λ_j for próximo de zero, haverá uma forte dependência linear entre as variáveis explicativas e os autovalores associados $\mathbf{t}_j$ determinarão o tipo de dependência linear.

Uma forma de se perceber esse fato, qual seja, o de se detectar multicolinearidade com os autovalores $\lambda_1,...,\lambda_p$, é examinando o número condicional de $(\mathbf{X}'\mathbf{X})$, definido como:

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

Geralmente, se $\kappa < 100$, não há problema sério com a multicolinearidade. Se $100 < \kappa < 1000$, existe sensível multicolinearidade. Caso $\kappa > 1000$, forte multicolinearidade é evidenciada.

A Tabela 2.1 apresenta os valores de FIV, o menor autovalor e correspondente autovetor da matriz de correlação entre as variáveis explicativas para as 49 observações e após a exclusão de duas delas, correspondendo aos países Hong Kong e Singapura, (pontos B e C da Figura 2.1).

Considerando as 49 observações, o menor autovalor da matriz de correlação das variáveis explicativas foi 0,0267, e ao analisar tanto o autovetor associado quanto os valores de FIV, podemos perceber a presença de multicolinearidade entre as variáveis explicativas DENS e AAC.

Os autores reforçam ainda a multicolinearidade detectada pela constatação da forte correlação linear entre DENS e AAC, com coeficiente de correlação linear de Pearson $r = 0{,}972$.

Ao excluir Hong Kong e Singapura da base de dados e novamente calcular o menor autovalor da matriz de correlação das variáveis explicativas, obteve-se 0,1802, além de uma queda significativa tanto dos FIV's das variáveis DENS e AAC, todos abaixo de 5, quanto das componentes do autovetor associado.

Além disso, a retirada desses pontos reduziu o coeficiente de correlação linear de Pearson entre DENS e AAC para 0,783.

Tabela 2.1 Comparação de Indicadores de Multicolinearidade

	Dados Completos (n = 49) $(\lambda_{min} = 0{,}0267)$		Hong Kong e Singapura excluídos (n = 47) $(\lambda_{min} = 0{,}1802)$	
Variáveis explicativas	v_1	FIV	v_1	FIV
TMI	0,0066	1,89	0,0991	1,94
RMP	0,0340	2,70	0,5373	2,75
DENS	0,7090	19,10	0,4386	2,69
AAC	-0,7034	18,85	-0,3627	2,68
MA	0,0275	3,49	0,6142	3,42
IES	0,0251	1,25	-0,0169	1,24

Fonte: Mason e Gunst (1985).

A Tabela 2.2 exibe os valores das alavancas h_{ii} e os resíduos "externamente studentizados" t_i de algumas observações selecionadas pelos próprios autores.

Apesar de nenhum resíduo studentizado ser exageradamente grande, vários dos h_{ii} são superiores a $2(p+1)/n = 0{,}286$.

Outro fato que merece observação é que os resíduos studentizados t_i não identificaram Hong Kong ou Singapura como observações influentes e, segundo os autores, outras medidas também falharam nesse sentido. Uma inspeção do diagrama de dispersão da Figura 2.1 indica que tal fato ocorreu devido a um possível efeito de

mascaramento, já que esses elementos são os pontos B e C, bastante próximos entre si. No entanto, os valores de h_{ii} e o diagrama de dispersão apontam para a possibilidade de sua influência conjunta no ajuste.

Tabela 2.2 - Medidas de Diagnósticos das Observações Selecionadas

Observação	h_{ii}	Resíduo studentizado t_i
Barbados	0,238	-2,026
Bélgica	0,043	1,209
Canadá	0,042	2,011
Estados Unidos	0,490	0,804
Hong Kong	0,511	-0,107
Índia	0,558	1,337
Japão	0,049	-2,799
Luxemburgo	0,084	2,356
Malta	0,688	1,506
Singapura	0,632	0,562
Tailândia	0,178	-2,402

Fonte: Mason e Gunst (1985).

A Tabela 2.3 diz apresenta a comparação entre as estimativas dos coeficientes de regressão padronizados para o grupo todo e sem as observações Hong Kong e Singapura. Fornece ainda os valores da estatística associada ao teste $H_0 : \beta_j = 0$, $t_j = \dfrac{\hat{\beta}_j}{DP\left(\hat{\beta}_j\right)}$.

Nessa tabela, é importante notar o fato de que quando Hong Kong e Singapura estão na base de dados ($n = 49$), os coeficientes de regressão das variáveis DENS e AAC não são estatisticamente significantes. A partir do momento que os dois países são retirados da amostra, os sinais dos dois coeficientes mudam. Além disso, o coeficiente da variável AAC passa a ser estatisticamente significante.

Tabela 2.3 - Estimativas dos Coeficientes da Regressão e Valores da Estatística t				
Variáveis explicativas	Estimativas dos Coeficientes		Valores da Estatística t	
	$n = 49$	$n = 47$	$n = 49$	$n = 47$
TMI	-1,870	-2,076	-3,31	-3,81
RMP	0,171	0,335	0,25	0,52
DENS	-1,094	0,622	-0,61	0,97
AAC	0,862	-1,447	0,48	-2,26
MA	2,298	2,204	2,99	3,05
IES	1,454	1,396	3,17	3,21

Fonte: Mason e Gunst (1985).

Mason e Gunst (1985) destacam que, de modo geral, quando pontos de alta alavanca aparecem em grupos, os valores individuais dos h_{ii} podem não necessariamente exceder $\frac{2(p+1)}{n}$, mesmo que esses pontos estejam distorcendo as estimativas de mínimos quadrados.

Outra possibilidade a considerar é que pontos discrepantes que ocorrem em três ou mais variáveis explicativas podem não ser facilmente perceptíveis em gráficos bidimensionais.

Por esse motivo, cinco procedimentos foram propostos pelos autores, com o objetivo de auxiliar no diagnóstico de multicolinearidade causada especificamente por pontos discrepantes:

1°) Determinar se existe multicolinearidade no conjunto de dados. Para tal, faça uso de medidas de correlação, cálculo de autovalores e autovetores, fatores de inflação de variância e outros conceitos úteis.

2°) Identificar pontos de alavanca. O uso dos valores de h_{ii} é eficiente caso os pontos discrepantes não estejam em grupos. Caso, pela natureza do problema, se desconfie que os pontos discrepantes estejam reunidos em grupos, deve-se utilizar um ou mais dos procedimentos de diagnóstico propostos para detectar outliers em grupo (ver também o 4° procedimento).

3°) Se a multicolinearidade ocorre em pares de variáveis, construir diagramas de dispersão para determinar se os pontos discrepantes estão induzindo multicolinearidade.

4°) Construir gráficos de pares de componentes principais normalizados, definidos como $\mathbf{m}_j = l_j^{-1}\mathbf{Z}\mathbf{v}_j$ correspondentes aos maiores autovalores de $\mathbf{Z}'\mathbf{Z}$. Se pontos discrepantes se reúnem em grupos, gráficos pareados de componentes principais normalizados correspondentes aos maiores autovalores de $\mathbf{Z}'\mathbf{Z}$ podem ser úteis como forma de detectá-los. Hocking (1984) exibe um exemplo em que dois pontos de alta alavanca próximos são detectados através do gráfico de $\mathbf{m}_1 \times \mathbf{m}_3$, $\mathbf{m}_1$ e $\mathbf{m}_3$ respectivamente valores das 1ª e 3ª componentes principais calculadas para os dados. Esses gráficos também podem ser úteis como diagnósticos de pontos discrepantes de dimensões de ordem maior ou igual a três, caso os maiores valores em $\mathbf{v}_j$ sejam os mesmos que aqueles do autovetor que identifica a multicolinearidade.

5°) Eliminar do conjunto de dados as observações suspeitas de induzir a multicolinearidade. Caso observações sejam eliminadas, refazer os passos anteriores com o conjunto reduzido de dados. A remoção de outliers deve resultar na eliminação da multicolinearidade.

De acordo com os autores, se o conjunto de dados não possui multicolinearidade ou pontos de alavanca, gráficos de dispersão dos componentes principais deverão exibir uma distribuição aleatória das observações. Mas, se a multicolinearidade foi detectada, gráficos de dispersão de componentes principais correspondentes aos maiores autovalores de $\mathbf{Z}'\mathbf{Z}$ serão eficientes para identificar pontos de alavancagem agrupados ou pontos extremos em duas ou mais dimensões.

Os autovetores associados aos dois maiores autovalores dos dados do PNB são

$\mathbf{v}_5' = (-0{,}356, -0{,}187;\ 0{,}636; 0{,}637; 0{,}108; 0{,}126)$ e

$\mathbf{v}_6' = (0{,}410; 0{,}508; 0{,}273; 0{,}269; -0{,}552; -0{,}350)$.

É possível verificar que no caso de $\mathbf{v}_5'$, os valores das posições de número três e quatro (DENS e AAC) são altos e quase iguais, e que no caso de $\mathbf{v}_6'$, os valores das mesmas posições não são tão altos, mas apresentam os mesmos sinais e são de magnitudes parecidas. No que diz respeito ao autovetor $\mathbf{v}_1$ da Tabela 2.1, as posições de número três e quatro também apresentam valores elevados, são também de magnitudes semelhantes, apesar de sinais opostos.

É possível perceber na Figura 2.2 que Hong Kong e Singapura são pontos de alavanca. Isso pode ser detectado pelo exame dos componentes de $\mathbf{m}_5$, mas um diagrama de dispersão dos dois componentes principais é ainda mais eficaz. Como os componentes principais correspondentes aos maiores autovalores encontram-se na direção de maior variabilidade dos dados, eles são graficamente mais eficientes para identificar múltiplos outliers e também para determinar se pontos de alavanca são devidos a valores extremos em três ou mais dimensões. Assim, segundo os autores, a chave para saber se pontos de alavanca induzem multicolinearidade está no padrão de valores altos e de magnitudes semelhantes nos autovetores associados aos dois maiores autovalores de $\mathbf{Z}'\mathbf{Z}$ nas posições das variáveis que identificam a multicolinearidade.

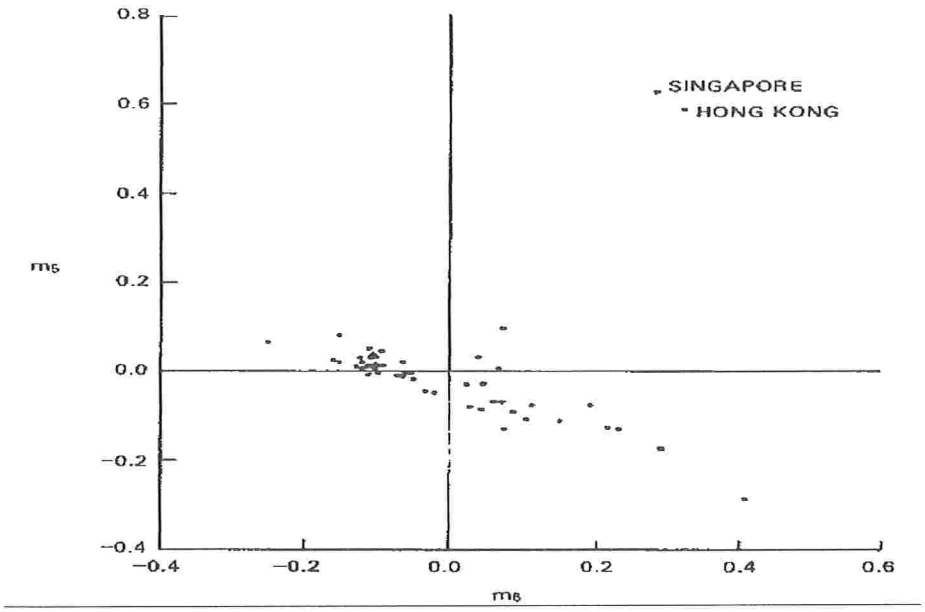


Figura 2.2 – Componentes Principais Normalizados para os Dois Maiores Autovalores.
Fonte: Mason e Gunst (1985).

Detectado o problema, haveria um amplo conjunto de soluções, desde a eliminação de observações até o uso de estimadores robustos, assunto esse que será discutido no Capítulo 6.

Capítulo 3

Medidas de Influência na Regressão em Cristas

3.1 - Introdução

No capítulo anterior foi visto que o estimador em cristas surgiu como uma possível forma de contornar o problema da multicolinearidade. Além disso, foi analisado o trabalho de Mason e Gunst (1985) que traz como resultado principal o fato de que algumas observações podem gerar multicolinearidade.

Nesse capítulo será apresentada uma situação completamente diferente. Serão analisados os efeitos que a multicolinearidade exerce na influência das observações por

meio do artigo de Walker e Birch (1988). Os autores propõem inicialmente medidas de influência apropriadas para o caso em que se utiliza o estimador em cristas. De posse dessas medidas, é apresentada uma análise comparativa da influência exercida por cada observação quando são adotados os dois procedimentos de estimação, mínimos quadrados e estimação em cristas.

O modelo de regressão linear utilizado é definido por

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon},$$

onde $\mathbf{y}$ é um vetor de variáveis aleatórias observáveis, $\mathbf{1}$ é um vetor contendo o valor 1 em todas as posições, β_0 é um parâmetro desconhecido, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ é uma matriz $n \times p$ centralizada e padronizada de constantes conhecidas $(\mathbf{1}'\mathbf{x}_i = 0,\ \mathbf{x}_i'\mathbf{x}_i = 1,\ i = 1, ..., p)$, $\boldsymbol{\beta}_1$ é um vetor de parâmetros desconhecidos e $\boldsymbol{\varepsilon}$ é um vetor de erros não observáveis com $E(\boldsymbol{\varepsilon}) = 0$ e $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.

Se $\mathbf{Z} = (\mathbf{1}, \mathbf{X})$, então o estimador de mínimos quadrados de $\boldsymbol{\beta}$ $[\boldsymbol{\beta}' = (\beta_0, \boldsymbol{\beta}_1')]$ é $\mathbf{b} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$ e o vetor de respostas ajustadas é $\hat{\mathbf{y}} = \mathbf{Z}\mathbf{b}$. O estimador de σ^2 é $s^2 = \mathbf{e}'\mathbf{e}/(n - p - 1)$, sendo que $\mathbf{e}$ é o vetor de resíduos $(\mathbf{y} - \hat{\mathbf{y}})$.

De acordo com os autores, a influência de uma observação pode ser vista como produto de dois fatores: uma função do resíduo e uma função da posição do ponto no espaço $\mathbf{Z}$. A posição ou alavancagem do *i-ésimo* ponto, como já visto, é medida por h_{ii}, o *i-ésimo* elemento da diagonal da matriz "hat" $\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$.

Entre as medidas de influência mais conhecidas destaca-se a medida

$$DFFITS\ (i) = \mathbf{z}_i(\mathbf{b} - \mathbf{b}(i))/\,E(\mathbf{z}_i, \mathbf{b})\ , \qquad (3.1)$$

sendo que $\mathbf{b}(i)$ é o estimador de mínimos quadrados de $\boldsymbol{\beta}$ sem a *i-ésima* observação e $E(\mathbf{z}_i, \mathbf{b})$ é um estimador do desvio padrão do valor ajustado. Portanto, $DFFITS(i)$

avalia a alteração no valor ajustado quando uma observação é eliminada. Assim, pode ser considerada como uma medida de influência de observações individuais.

Verifica-se que a medida (3.1) pode ser escrita como um produto de dois fatores, um dependendo do resíduo e o outro dependendo do valor de alavancagem. Assim,

$$DFFITS\,(i) = [e_i / s(i)]\left[h_{ii}^{1/2} / (1 - h_{ii})\right], \tag{3.2}$$

sendo que e_i é o *i-ésimo* resíduo, e $s(i)$ é o estimador de mínimos quadrados de σ quando a *i-ésima* observação é eliminada.

Outra importante medida de influência, já vista anteriormente, é D de Cook, definida por

$$D_i = (\mathbf{b} - \mathbf{b}(i))' \mathbf{Z}'\mathbf{Z} (\mathbf{b} - \mathbf{b}(i)) / ((p+1)s^2), \tag{3.3}$$

em que $s^2 = \hat{\sigma}^2$ é o quadrado médio do resíduo do modelo ajustado com todas as observações, e que pode ser escrito na forma

$$D_i = (e_i^2 / (p+1)s^2)[h_i / (1 - h_i)^2]. \tag{3.4}$$

Para detectar pontos influentes, alguns autores sugerem que D_i deve ser comparado com quantis da distribuição $F(p+1, n-p-1)$.

Ressalta-se, novamente, que essas medidas são úteis para detectar influência de observações individuais. Essas medidas sofrem, no entanto, de um grave problema, a saber, o de *mascarar* a influência potencial de outras observações, assunto esse já abordado na seção anterior.

3.2 – Medidas de Influência em Regressão em Cristas

Mason e Gunst (1985) procuraram analisar os efeitos que algumas observações discrepantes exerciam na multicolinearidade.

Já no caso da situação contrária, ou seja, sobre os efeitos que a multicolinearidade pode exercer nas medidas de influência, pouco se sabe. Belsley et al. (1980, p. 210) utilizando estimadores de regressão em cristas com o objetivo de reduzir os efeitos da multicolinearidade, percebeu que a maioria das medidas de influência das observações era menor que as correspondentes medidas calculadas no contexto de mínimos quadrados. No entanto, verificou-se que a influência de algumas observações chegou a aumentar mesmo quando a multicolinearidade foi controlada. Com base nesses fatos, os autores afirmaram que como a multicolinearidade pode disfarçar pontos discrepantes, então a sua redução deveria ser o primeiro passo para a detecção efetiva de dados incomuns.

Com relação ao estimador em cristas,

$$\mathbf{b}^* = \left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^*\right)^{-1}\mathbf{Z}'\mathbf{y}, \tag{3.5}$$

onde $\mathbf{I}^* = diag(0, 1, ..., 1)$ de dimensão $p+1$, Marquardt (1970) demonstrou que o estimador (3.5) também pode ser obtido por mínimos quadrados se os dados fossem aumentados como

$$\mathbf{y}_A = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}, \qquad \mathbf{X}_A = \begin{bmatrix} \mathbf{X} \\ k^{1/2}\mathbf{I} \end{bmatrix}, \tag{3.6}$$

onde $\mathbf{0}$ é um vetor $(p+1)-$dimensional de zeros e $\mathbf{I}$ é a matriz identidade de ordem $p+1$.

Ao utilizarmos o estimador (3.5), o vetor de valores ajustados será

$$\begin{aligned} \hat{\mathbf{y}}^* &= \mathbf{Z}\mathbf{b}^* \\ &= \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^*\right)^{-1}\mathbf{Z}'\mathbf{y}. \end{aligned}$$

Portanto, a matriz $\mathbf{H}^* = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^*\right)^{-1}\mathbf{Z}'$ assume uma função similar à da matriz "hat" na estimação por mínimos quadrados. O *i-ésimo* valor previsto pode ser escrito em termos dos elementos de $\mathbf{H}^*$ como

$$\hat{y}_i^* = \sum_{j=1}^{n} h_{ij}^* y_j .$$

Consequentemente, $\partial \hat{y}_i^* / \partial y_i = h_{ii}^* \equiv h_i^*$ e com isso, os elementos da diagonal da matriz "hat" do estimador em cristas podem ser interpretados, assim como no caso de mínimos quadrados, como um valor de alavancagem. Vale lembrar, contudo, que a matriz $\mathbf{H}^*$ não é uma matriz projeção, pois não é idempotente.

A técnica de Decomposição de Valor Singular (DVS) ou decomposição espectral permite que $\mathbf{Z}$ seja escrita como $\mathbf{Z} = \mathbf{UDV}'$, onde $\mathbf{D}$ é uma matriz diagonal $(p+1)\times(p+1)$ cujo *i-ésimo* elemento da diagonal é $\lambda_i^{1/2}$, sendo λ_i o autovalor de $\mathbf{Z}'\mathbf{Z}$, as colunas de $\mathbf{V}$ são os autovetores de $\mathbf{Z}'\mathbf{Z}$ e o elemento de ordem ij da matriz $\mathbf{U}_{(u_{ij})}$ $n\times(p+1)$ é tal que $u_{ij}\lambda_j^{1/2}$ é a projeção da *i-ésima* linha, $\mathbf{z}_i$, no *j-ésimo* autovetor de $\mathbf{Z}$. Usando a DVS, Lichtenstein e Velleman (1983), citado em Walker e Birch (1988), chegaram ao valor de alavancagem do *i-ésimo* ponto na regressão em cristas por meio da expressão

$$h_i^* = \sum_{j=1}^{p+1} \frac{\lambda_j}{\lambda_j + k} u_{ij}^2 .$$

Dois fatos importantes derivam desta expressão. O primeiro é que para $k > 0$, $h_i^* < h_i$ para $i = 1,...,n$. Isso significa que para qualquer observação, o valor de alavancagem em cristas será sempre menor do que seu correspondente em mínimos quadrados. O segundo é que o valor de alavancagem decai monotonicamente à medida que k aumenta.

Especificamente o segundo fato mencionado pode dar a entender que a influência de uma observação é modificada apenas, e tão somente, à medida que k aumenta. Dessa forma, é importante relembrar que, conforme destacado no início do

capítulo, a influência de uma observação é produto de dois fatores atuando conjuntamente: (i) resíduos e (ii) valores de alavancagem.

Sobre o fator resíduos, os autores definiram resíduo em cristas como:

$$e_i^* = y_i - \hat{y}_i^* = y_i - \mathbf{z}_i\mathbf{b}^*,$$

o qual, usando a técnica DVS, Tripp (1983, p.87), citado em Walker e Birch (1988), reescreveu como

$$e_i^* = e_i + k\sum_{j=1}^{n} y_j \left[\sum_{m=1}^{p} \frac{u_{im}u_{jm}}{\lambda_m + k}\right]. \tag{3.7}$$

Os autores observam que a segunda parcela em (3.7) pode ser tanto negativa quanto positiva. Assim, o valor absoluto do resíduo em cristas para uma dada observação pode ser maior ou menor que seu correspondente no caso de mínimos quadrados.

Uma versão alternativa para a medida de influência *DFFITS* no contexto de regressão em cristas é descrita, pelos autores, por

$$DFFITS^*(i) = \mathbf{z}_i\left(\mathbf{b}^* - \mathbf{b}^*(i)\right) / E\left(\mathbf{z}_i\,\mathbf{b}^*\right),$$

na qual $\mathbf{b}^*(i)$ é o estimador em cristas (3.5) calculado sem a *i-ésima* observação e o denominador é um estimador do erro padrão do valor ajustado pela regressão em cristas. Se k for não estocástico, então:

$$\begin{aligned} SE\left(\mathbf{z}_i\,\mathbf{b}^*\right) &= s\left[\mathbf{z}_i\left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^*\right)^{-1}\mathbf{Z}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^*\right)^{-1}\mathbf{z}'_i\right]^{1/2} \\ &= s\left[\sum_{j=1}^{n} h_{ij}^{*2}\right]^{1/2}. \end{aligned}$$

Assim, $DFFITS^*(i)$ poderá ser reescrita como:

$$DFFITS^*(i) = \frac{\mathbf{z}_i\left(\mathbf{b}^* - \mathbf{b}^*(i)\right)}{s(i)\left[\sum_{j=1}^{n} h_{ij}^{*2}\right]^{1/2}}. \tag{3.8}$$

Uma versão alternativa para a distância de Cook D_i, adaptada também ao contexto de regressão em cristas pode ser dada pela expressão

$$D_i^* = \left(1/\left((p+1)s^2\right)\right)\left(\mathbf{b}^* - \mathbf{b}^*(i)\right)' \mathbf{Z}'\mathbf{Z}\left(\mathbf{b}^* - \mathbf{b}^*(i)\right). \qquad (3.9.1)$$

A medida D_i^* também pode ser escrito como

$$D_i^* = \left(1/(p+1)s^2\right)\left(\hat{\mathbf{y}}^* - \hat{\mathbf{y}}^*(i)\right)'\left(\hat{\mathbf{y}}^* - \hat{\mathbf{y}}^*(i)\right),$$

sendo que $\hat{\mathbf{y}}^* = \mathbf{Z}\mathbf{b}^*(i)$.

Uma segunda versão alternativa para D_i é dada por

$$D_i^{**} = \left(1/\left(ps^2\right)\right)\left(\mathbf{b}^* - \mathbf{b}^*(i)\right)'\left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^*\right)\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^*\right)\left(\mathbf{b}^* - \mathbf{b}^*(i)\right). \qquad (3.9\text{-}2)$$

É possível verificar que tanto D_i^* quanto D_i^{**} se reduzem à D_i quando $k = 0$.

Baseado no critério (3.6) das matrizes aumentadas, Belsley et al. (1980, p. 208) sugeriu $2[(p+1)/(n+p+1)]^{1/2}$ como ponto de corte para a medida $DFFITS^*(i)$ quantis da distribuição $F(p+1, n-p-1)$.

Foram ainda obtidas versões aproximadas dessas medidas de influência, escritas em função dos valores de alavancas e dos resíduos, que serão apresentadas no final desse capítulo.

3.3 – Exemplos e Resultados

Walker e Birch (1988) analisaram duas bases de dados previamente existentes: a primeira encontra-se em Longley (1967) e a segunda foi extraída de Hill (1977).

Com relação à primeira base de dados, o número condicional κ é 43.275, indicando severa multicolinearidade. Cook (1977) calculou os valores de D_i para esses

dados e notou que as observações 5, 16, 4, 10 e 15 (nessa exata ordem) eram as mais influentes, no contexto de mínimos quadrados.

Walker e Birch (1988) calcularam, por outro lado, os valores de D_i^* para esses dados, utilizando os estimadores de mínimos quadrados e em cristas. Os resultados estão organizados na Tabela 3.1.

Tabela 3.1 – Observações mais Influentes de acordo com D_i^*: Dados de Longley (1967)

$k = 0$ (Mínimos Quadrados)			$k = 0{,}0002$	
Observação	D_i^*		Observação	D_i^*
5	0,614		16	0,582
16	0,467		10	0,251
4	0,244		4	0,219
10	0,235		15	0,145
15	0,170		1	0,142

Fonte: Walker e Birch (1988).

É importante lembrarmos que há vários métodos para a obtenção de k. O método escolhido pelos autores consiste em escolher o valor de k que minimiza

$$C_k = \left(SQR_k / s^2\right) - n + 2 \cdot tr\left(\mathbf{H}^*\right),$$

onde SQR_k é a soma dos quadrados dos resíduos da regressão em cristas e s^2 é o estimador de mínimos quadrados para σ^2. Ao utilizar esse método, os autores obtiveram, para os dados de Longley, o valor $k = 0{,}0002$.

Inicialmente, podemos perceber que das cinco observações influentes determinadas quando utilizado o método dos mínimos quadrados, exatamente quatro delas foram consideradas influentes quando adotado o estimador em cristas. Além disso, é possível perceber, pela Tabela 3.1, que a influência dessas observações, em média, diminui quando é adotado o estimador em cristas.

Podemos verificar, ainda, na Tabela 3.1 que as observações mais influentes calculadas por um método já são bem diferentes quando comparadas com o outro

método: as medidas de influência, em geral, decrescem quando utilizado o estimador em cristas, com a ressalva de que a observação 5 deixou de estar entre as mais influentes e a influência das observações 10 e 16 aumentou.

As medidas de influência $DFFITS*(i)$ e D_i^* foram calculadas para vários valores de k e os resultados para as cinco observações mais influentes no ajuste de mínimos quadrados podem ser vistas nas Figuras 3.1 e 3.2. É importante ressaltar que a curva referente à observação 5 não aparece nessas figuras.

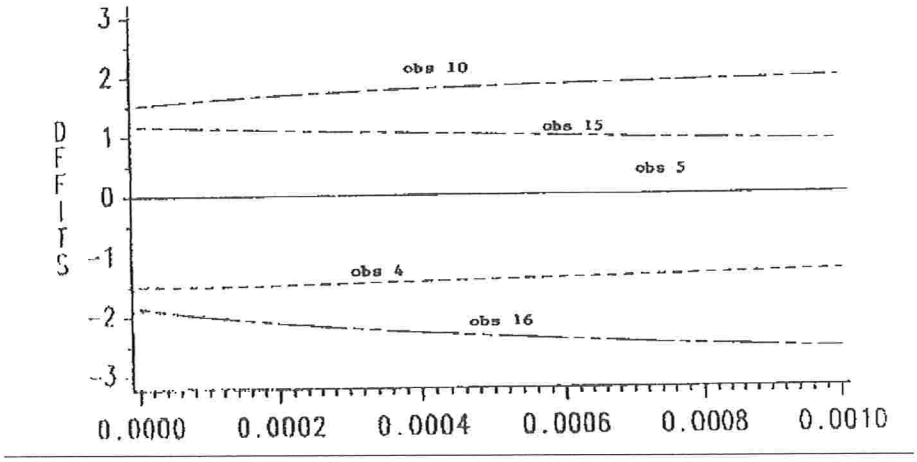


Figura 3.1 – DFFITS* versus k – dados de Longley
Fonte: Walker e Birch (1988).

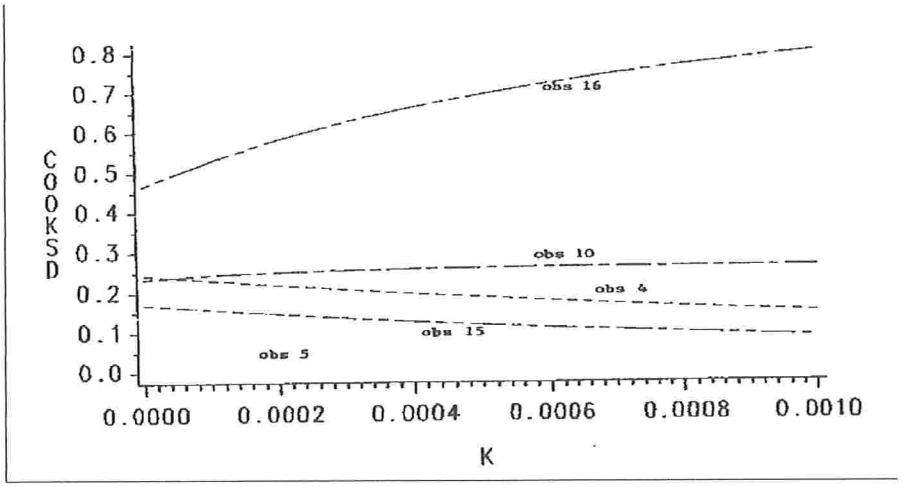


Figura 3.2 – D* versus k - dados de Longley

Fonte: Walker e Birch (1988).

Ao analisar o comportamento das duas medidas de influência em função de k, podemos perceber que, em ambos os casos, as influências se mantêm relativamente constantes, exceto a influência da observação 16 que, no caso de D_i^*, aumenta à medida que k cresce.

Os autores perceberam, ainda, que a influência da observação 5 decresce tanto em $DFFITS^*(i)$ quanto em D_i^*, apesar de ser a mais influente em mínimos quadrados.

Com relação ao comportamento dos valores das alavancas, descrito na Figura 3.3, os autores notaram que todos os valores decrescem em função de k, sendo que os correspondentes à observação 5 decrescem mais rapidamente que os demais e os valores da observação 16 são os mais altos. Ressalta-se, novamente, que embora a observação 5 esteja destacada no gráfico, os valores de suas medidas de influência não se encontram na figura presente no artigo.

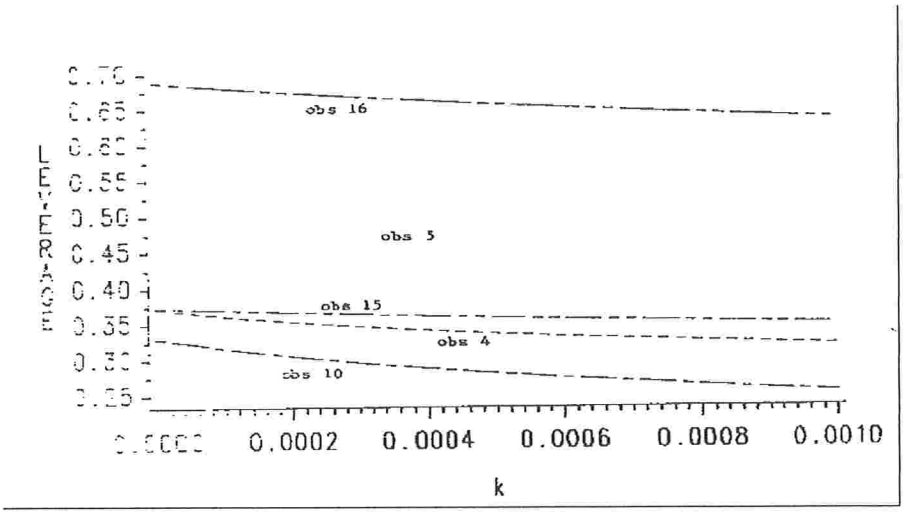


Figura 3.3 – Alavanca versus k - dados de Longley
Fonte: Walker e Birch (1988).

A Figura 3.4 descreve o comportamento dos resíduos em função de k. Segundo os autores, o resíduo da observação 5 decresce significativamente, fato que, novamente, não é possível averiguar na figura. Por outro lado, os resíduos das observações 10 e 16 aumentaram em valor absoluto.

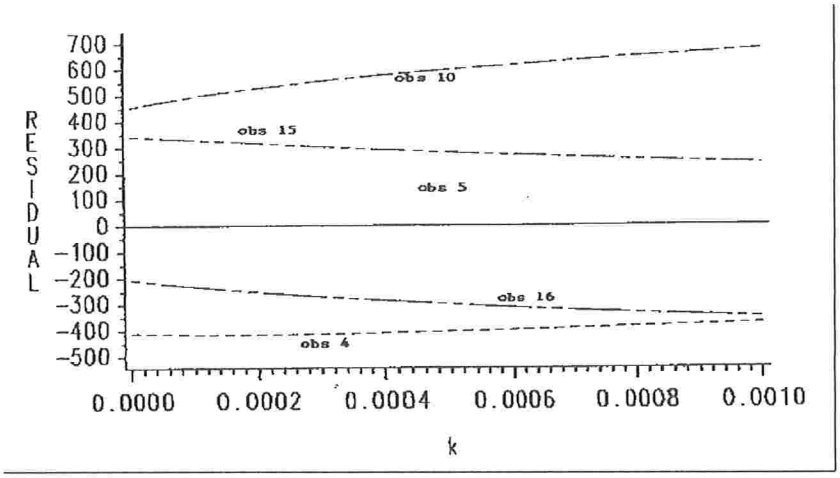


Figura 3.4 – Resíduos versus k - dados de Longley
Fonte: Walker e Birch (1988).

Com relação à base de dados de Hill (1977), o número condicional κ é 57,14, que, por ser menor que 100, indica um grau moderado para a multicolinearidade.

Inicialmente, podemos perceber pela Tabela 3.2 que, das cinco observações influentes obtidas no ajuste pelo método dos mínimos quadrados ($k = 0$), todas foram consideradas influentes pelo estimador em cristas.

Tabela 3.2 – Observações mais influentes de acordo com *DFFITS* *: Dados de Hill (1977)

$k = 0$ (Mínimos Quadrados)			$k = 0{,}03$	
Observação	*DFFITS* *		Observação	*DFFITS* *
2	7,9825		8	2,0099
1	-4,0852		1	-1,9903
8	1,8530		2	1,7687
15	-1,2724		15	-1,5024
12	1,1957		12	1,4437

Fonte: Walker e Birch (1988).

Para os dados de Hill, o valor de k que minimiza C_k é 0,03. A Tabela 3.2 apresenta os valores da medida $DFFITS*(i)$ quando $k=0$ e $k=0{,}03$, para as cinco observações mais influentes. É possível perceber, pela Tabela 3.2, que a influência dessas observações, em média, diminui no procedimento de regressão em cristas.

Segundo os autores, para as cinco observações mais influentes, os valores da medida $DFFITS*(i)$ das observações 1 e 2 decrescem (em valor absoluto) rapidamente quando k aumenta, enquanto que as demais se mantiveram relativamente estáveis. Isso pode ser percebido pela Figura 3.5, com a ressalva de que a curva associada à observação 2 não apareceu.

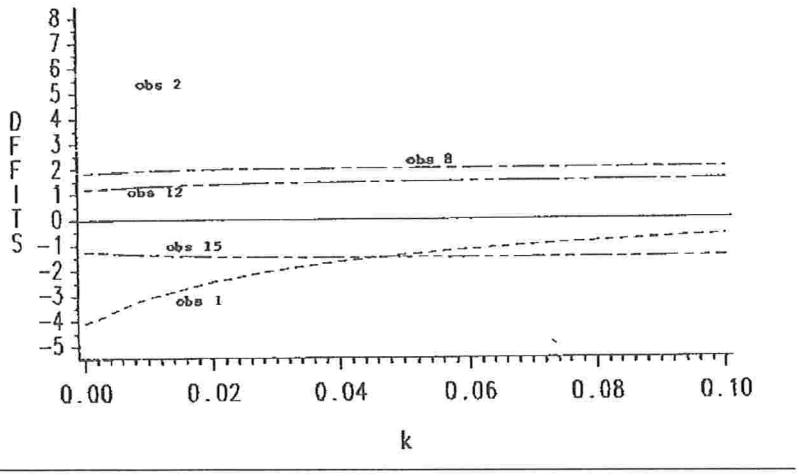


Figura 3.5 – DFFITS* versus k - dados de Hill
Fonte: Walker e Birch (1988).

Ainda segundo os autores, o mesmo foi verificado para a medida D_i^*, cuja figura não foi apresentada.

Verificou-se, ainda, que os valores das alavancas dessas cinco observações mais influentes decrescem com k a, aproximadamente, a mesma velocidade.

A Figura 3.6 apresenta o comportamento dos resíduos em função de k. Nela, os autores perceberam que enquanto os resíduos das observações 1 e 2 decresceram rapidamente (em valor absoluto), os resíduos das demais observações cresceram significativamente. Novamente deve-se ressaltar que a observação 2 não está presente na figura.

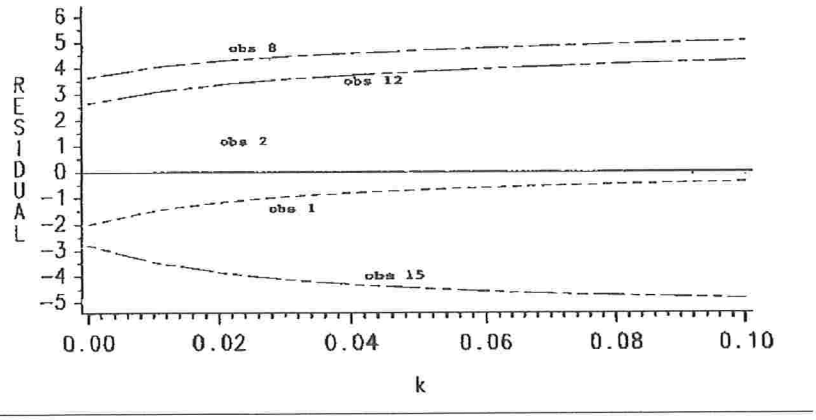


Figura 3.6 – Resíduos versus k - dados de Hill
Fonte: Walker e Birch (1988).

Finalizando o artigo, os autores fornecem aproximações para as medidas de influência propostas. Consecutivamente, essas medidas aproximadas foram calculadas para os conjuntos de dados discutidos anteriormente e os principais resultados foram analisados.

Segundo os autores, para pequenos valores de k e ou observações com baixa alavanca, medidas aproximadas de influência podem ser obtidas por meio do teorema de Sherman, Morrison e Woodbury (Belsley et al. 1980, pág. 64).

O resultado principal é expresso por

$$\mathbf{b}^*(i) \approx \mathbf{b}^* - \left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^*\right)^{-1} \mathbf{z}'_i \, e_i^* / \left(1 - h_i^*\right).$$

Com base nesse resultado, versões aproximadas de $DFFITS*(i)$ e D_i^* podem ser obtidos como

$$DFFITS_a^*(i) = \frac{e_i^*}{s(i)} \frac{h_i^*}{1 - h_i^*} \left[\frac{1}{\sum_j h_{ij}^{*2}} \right]^{1/2}$$

(3.10)

e

$$D_a^*(i) = \left(e_i^{*2} / (ps^2)\right) \left[\sum_j h_{ij}^{*2} / \left(1 - h_i^*\right)^2\right], \qquad (3.11)$$

onde s^2 e $s(i)$ são calculados por mínimos quadrados. É possível verificar que, para $k = 0$, (3.10) se reduz à expressão $DFFITS(i)$ e (3.11) à expressão D_i, ambos no contexto de mínimos quadrados, desde que $\sum_j h_{ij}^{*2} = h_i$.

Dessa maneira, as duas medidas de influência estão escritas como função dos fatores alavancagem e resíduos e, também, estão adaptadas para a regressão em cristas.

As expressões (3.10) e (3.11) foram utilizadas pelos autores para analisar as duas bases de dados.

Para os dados de Longley, os gráficos de (3.10) e (3.11) em função de k (não apresentados pelos autores) são virtualmente indistinguíveis quando comparados com as Figuras 3.1 e 3.2.

Já, para os dados de Hill, as medidas aproximadas não foram tão precisas. A precisão é particularmente ruim para as observações 1 e 2, que possuem altos valores de alavanca ($h_1 = 0{,}8369$ e $h_2 = 0{,}9218$, respectivamente).

Segundo os autores, tal fato não é uma simples coincidência. Essas observações que apresentam grandes diferenças entre as medidas exata e aproximada são aqueles pontos que sofreriam mais os efeitos de centralização e escalonamento da matriz $\mathbf{Z}$.

A principal vantagem das medidas aproximadas é que, como no diagnóstico em mínimos quadrados, não seria necessário recalcular a estimativa de $\boldsymbol{\beta}$ a cada observação eliminada. Assim, para um dado valor de k, as medidas aproximadas (3.10) e (3.11) podem ser calculadas após um único ajuste do modelo de regressão em cristas.

Assim, foi mostrado nesse capítulo que quando o estimador em cristas é utilizado para reduzir o efeito da multicolinearidade, mudanças significativas podem ser observadas na influência de algumas observações.

Observamos, também, que medidas de influência baseadas no estimador em cristas, em média, são menores quando comparadas com aquelas obtidas pelo método de mínimos quadrados.

Após identificar e eliminar os casos mais influentes, um novo modelo deve ser ajustado e tanto multicolinearidade quanto medidas de influência devem ser reexaminadas.

Medidas aproximadas de influência foram apresentadas com o objetivo de simplificar os cálculos. A vantagem dessas medidas reside no fato de que, com elas, não é mais necessário reescalonar a matriz $\mathbf{Z}(i)$, ou seja, torná-la uma matriz com colunas de comprimento unitário, para reestimar $\mathbf{b}^*(i)$ toda vez que uma linha é eliminada.

Mas, segundo Walker e Birch (1988), tais medidas não geram boas aproximações em pontos de alavancas elevadas e, nesse caso, o uso de medidas exatas é mais recomendado.

Capítulo 4

Análise de Influência Local na Regressão em Cristas

4.1 - Introdução

Nesse capítulo serão apresentadas duas medidas de influência local em regressão em cristas, obtidas por Billor e Loynes (1999).

O método da influência local foi desenvolvido originalmente por Cook (1986) e é aplicável apenas em procedimentos de estimação via função de verossimilhança.

Para o cálculo das medidas de influência, de acordo com essa abordagem, Billor e Loynes (1999) escreveram o estimador em cristas como um estimador de máxima

pseudo-verossimilhança. Os autores concluem a análise com um estudo comparativo entre as medidas propostas e posterior aplicação a um particular conjunto de dados.

4.2 – Medidas de Influência Local

Apresentaremos, a seguir, uma breve descrição das medidas de influência local propostas por Cook (1986).

Seja $L(\boldsymbol{\theta})$ o logaritmo da função de verossimilhança para um modelo inicial, sendo $\boldsymbol{\theta}$ um vetor $p\times 1$ de parâmetros desconhecidos com estimador de máxima verossimilhança dado por $\hat{\boldsymbol{\theta}}$. São introduzidos distúrbios no modelo através do vetor $\mathbf{w}$, $m\times 1$, $\mathbf{w}\subset\boldsymbol{\Omega}$, $\boldsymbol{\Omega}\subset\Re^m$, onde $\boldsymbol{\Omega}$ representa um conjunto aberto de possíveis pequenos distúrbios. Do ponto de vista prático, $\mathbf{w}$ refletiria qualquer esquema de perturbação. Como exemplo, Cook (1986) cita a introdução de pequenas modificações nas variáveis explicativas ou perturbações na matriz de covariância de modelos de regressão.

Nessas condições, consideremos $\mathcal{L}(\boldsymbol{\theta}\,|\,\mathbf{w})$ o logaritmo da função de verossimilhança que corresponde ao modelo que sofreu perturbação e $\hat{\boldsymbol{\theta}}_w$ o estimador de máxima verossimilhança correspondente a esse modelo. Suponha que exista um ponto $\mathbf{w}_0$ em $\boldsymbol{\Omega}$ que representa a ausência de perturbação nos dados, de modo que $L(\boldsymbol{\theta})=\mathcal{L}(\boldsymbol{\theta}\,|\,\mathbf{w}_0)$, e assuma que $\mathcal{L}(\boldsymbol{\theta}\,|\,\mathbf{w})$ seja duplamente diferenciável e contínua em uma vizinhança de $(\hat{\boldsymbol{\theta}}', \mathbf{w}'_0)$. O deslocamento de verossimilhanças de Cook é definido como

$$LD(\mathbf{w}) = 2\left[L(\hat{\boldsymbol{\theta}}) - L(\hat{\boldsymbol{\theta}}_w)\right],$$

que permitiria comparar as estimativas $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_w$, podendo, assim, avaliar a influência dos distúrbios $\mathbf{w}$. Grandes valores de $LD(\mathbf{w})$ indicam que $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_w$ diferem consideravelmente em relação ao contorno da função de verossimilhança sem perturbação $L(\boldsymbol{\theta})$.

Esse método é baseado no estudo do comportamento local de um gráfico de influência $\alpha(\mathbf{w}) = (\mathbf{w}', LD(\mathbf{w}))$ ao redor de $\mathbf{w}_0$. O procedimento consiste em considerar $\mathbf{w}$ como $\mathbf{w}(a) = \mathbf{w}_0 + a\mathbf{d}$, $a \in \Re$ e $\mathbf{d}$ um vetor direção de comprimento unitário. Cook (1986) sugere investigar a direção na qual a medida de influência $LD(\mathbf{w})$ muda localmente mais rapidamente, isto é, a curvatura máxima da superfície $\alpha(\mathbf{w})$. De acordo com Cook (1986, pág. 139), a curvatura máxima de LD é dada por

$$C_{\max} = \max_{\|d\|=1} 2|\mathbf{d}'\mathbf{F}\mathbf{d}|,$$

em que $\mathbf{F}$ é uma matriz $m \times m$ definida por

$$\mathbf{F} = \boldsymbol{\Delta}'\mathbf{Q}^{-1}\boldsymbol{\Delta},$$

$\boldsymbol{\Delta}$ é a matriz $p \times m$ $(p = \dim(\boldsymbol{\theta}), m = \dim(\mathbf{w}))$ com elementos

$$\Delta_{ij} = \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{w})}{\partial \boldsymbol{\theta}_i \partial \mathbf{w}_j},$$

avaliados em $\hat{\boldsymbol{\theta}}$ e $\mathbf{w}_0$, e $-\mathbf{Q}$ representa a matriz de informação observada do modelo sem distúrbios $\mathbf{Q} = [\partial^2 L(\boldsymbol{\theta}) \mid \partial\boldsymbol{\theta}_i \partial\boldsymbol{\theta}_j]$, avaliada em $\hat{\boldsymbol{\theta}}$. Verifica-se (ver, por exemplo, Morisson, 1976, pág. 73) que a maximização de $|\mathbf{d}'\mathbf{F}\mathbf{d}|$, sujeito à restrição que $\mathbf{d}'\mathbf{d} = 1$, resulta em $\mathbf{d}_{\max}$, que representa o autovetor correspondente ao maior autovalor absoluto $C_{\max}$ de $\mathbf{F}$. A direção do vetor $\mathbf{d}_{\max}$ seria aquela que produziria a maior mudança local

nas estimativas dos parâmetros. Dessa forma, o vetor $\mathbf{d}_{max}$ é utilizado para identificar as observações que podem ser bastante significativas na análise.

Cook (1986) sugere como referência geral uma curvatura igual a 2, sendo que curvaturas maiores que esse valor indicariam notável sensibilidade local.

Billor e Loynes (1999) acreditam que a medida LD de Cook pode gerar respostas inesperadas e, por esse motivo, propuseram uma medida diferente, descrita por

$$LD^*(\mathbf{w}) = -2\left[L(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\hat{\boldsymbol{\theta}}_w \mid \mathbf{w})\right].$$

A medida $LD^*(\mathbf{w})$ compararia, então, as funções de verossimilhança das duas situações consideradas, com e sem perturbação.

Assim, temos uma superfície $(\mathbf{w}', LD^*(\mathbf{w}))$ com linhas dispersas em várias direções. Se tomarmos $\mathbf{w} = \mathbf{w}_0 + a\mathbf{d}$, para $\mathbf{d}$ fixado enquanto a varia, então o comportamento de $LD^*(\mathbf{w})$ é de interesse real. Nesse caso, a primeira derivada avaliada em $\mathbf{w}_0$, com raras exceções, é diferente de zero e esta primeira derivada, vista como função de $\mathbf{d}$, proporciona informações úteis sobre o comportamento local de $LD^*(\mathbf{w})$. Em particular, a inclinação máxima l_{max} e a correspondente direção $\mathbf{d}_{max}$ são importantes quando $m \neq 1$. Por esse motivo, para $m \geq 2$, sugerem o uso da medida

$$l_{max} = \left|\nabla LD^*(\mathbf{w}_0)\right|,$$

onde $\nabla LD^*(\mathbf{w}_0)$ é o vetor gradiente da função LD^* em $\mathbf{w}_0$, ou seja, o vetor das derivadas parciais de LD^* calculadas em $\mathbf{w}_0$. Os autores acreditam que pelo fato de envolver apenas o cálculo da primeira derivada, essa medida acaba sendo mais fácil de ser obtida, quando comparada com a medida de Cook.

Observa-se que a inclinação máxima, escrita em função de L, é

$$l_{\max} = 2\left|\nabla \mathcal{L}\left(\hat{\theta} \mid \mathbf{w}\right)\right|.$$

4.3 – Estimador em Cristas de Máxima Pseudo-Verossimilhança

O estimador em cristas não é, por assim dizer, um estimador de máxima verossimilhança. Por outro lado, as medidas de influência local só podem ser utilizadas em procedimentos de estimação via função de verossimilhança. Por esse motivo, Billor e Loynes (1999) escreveram o estimador em cristas numa forma alternativa que seria obtida pelo método da máxima verossimilhança. Para tal fim, consideraram um modelo de regressão linear múltipla

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4.1}$$

onde $\mathbf{X}$ é uma matriz conhecida $n \times p$ padronizada, $\boldsymbol{\beta}$ é um vetor $p \times 1$ de parâmetros conhecidos, $\boldsymbol{\varepsilon}$ é o vetor de erros $p \times 1$ independentes e com distribuição normal com média zero e variância desconhecida σ^2. Admitiu-se adicionalmente que, nesse modelo, o termo constante não foi incluído.

Marquardt (1970) demonstrou que o estimador em cristas é equivalente ao estimador de mínimos quadrados quando os dados são suplementados por um conjunto de dados fictícios tomados de acordo com a matriz de planejamento ortogonal Hk e a variável resposta $\mathbf{Y}$ sendo zero em cada ponto fictício adicionado.

O modelo aumentado com matriz de planejamento $(n+p)\times p$

$$\mathbf{X}_a = \begin{pmatrix} \mathbf{X} \\ (k\mathbf{I})^{1/2} \end{pmatrix}$$

e o vetor $(n+p)\times 1$ de variáveis resposta $\mathbf{Y}_a' = (\mathbf{Y}'\, 0')$ pode ser escrito como

$$\mathbf{Y}_a = \mathbf{X}_a\boldsymbol{\beta} + \boldsymbol{\varepsilon}_a, \tag{4.2}$$

onde $\boldsymbol{\varepsilon}_a$ representa um vetor aleatório cujas componentes são variáveis aleatórias independentes e normalmente distribuídas com média zero e variância σ^2. A função densidade de $\mathbf{Y}_a$ será denominada função pseudo-densidade e a correspondente função de pseudo-verossimilhança será descrita por

$$L_p(\boldsymbol{\beta}) = \frac{n+p}{2}\log 2\pi - \frac{n+p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}'_i\boldsymbol{\beta})^2 + k\boldsymbol{\beta}'\boldsymbol{\beta}\right].$$

O estimador de máxima pseudo-verossimilhança é resultante de $\dfrac{\partial L_p(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 0$.

Como

$$\sum_{i=1}^{n}\left(\mathbf{y}_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 + k\boldsymbol{\beta}'\boldsymbol{\beta}$$

pode ser escrito na forma

$$\begin{aligned}&(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + k\boldsymbol{\beta}'\boldsymbol{\beta}\\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'[\mathbf{X}'\mathbf{X} + k\mathbf{I}]\boldsymbol{\beta},\end{aligned}$$

derivando-se essa expressão com relação a $\boldsymbol{\beta}$ e igualando a zero, obtem-se

$$2[\mathbf{X}'\mathbf{X} + k\mathbf{I}]\boldsymbol{\beta} - 2\mathbf{X}'\mathbf{y} = 0.$$

Resolvendo essa equação, em decorrência da utilização da matriz aumentada, obtêm-se a solução $\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$, que é o estimador em cristas. Uma vez que o estimador em cristas é o estimador de máxima pseudo-verossimilhança para o modelo considerado, a medida de influência local de Cook pode ser aplicada na regressão em cristas.

4.4 – Análise da Influência Local em Regressão em Cristas

Nesta seção, apresentaremos a análise de influência local e as correspondentes medidas de influência na regressão em cristas. Foram analisadas perturbações na função de variância.

Dessa forma, considere o modelo originalmente descrito em (4.1). Esse modelo supõe homogeneidade na variância do erro, ou seja, $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$. Supondo que pequenos distúrbios são introduzidos na variância de ε_i por meio de um vetor de distúrbios $\mathbf{w}$, $n \times 1$, a suposição relativa à variância dos erros passa a ser escrita de forma diferente, isto é, como $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{W}^{-1}$, onde σ^2 é considerado conhecido e $\mathbf{W} = diag(1+\mathbf{w}_1,...,1+\mathbf{w}_n)$ é uma matriz $n \times n$, diagonal. Dessa forma, $\text{var}(\varepsilon_i) = \sigma^2 (1+\mathbf{w}_i)^{-1}$. Apenas distúrbios da parte real do modelo (4.2) serão considerados, pois aqueles referentes à parte fictícia não são significativos.

Nessas condições, a função de pseudo-verossimilhança com distúrbios para o modelo aumentado é

$$\mathcal{L}_p(\boldsymbol{\beta} \mid \mathbf{w}) = cons\tan te - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}'_i\,\boldsymbol{\beta})^2 \mathbf{w}_i^* + k\boldsymbol{\beta}'\boldsymbol{\beta}\right] + \frac{1}{2}\sum_{i=1}^{n}\log \mathbf{w}_i^*,$$

onde $\mathbf{w}_i^* = 1 + \mathbf{w}_i$, $\mathbf{w}_i$ sendo o *i-ésimo* componente do vetor $n \times 1$ de distúrbios $\mathbf{w}$.

Para o cálculo da curvatura máxima $C_{\max}$ é necessária a obtenção dos componentes individuais da matriz da informação observada $-\mathbf{Q}$ e da matriz

$$\boldsymbol{\Delta} = \left[\frac{\partial^2 \mathcal{L}_p(\boldsymbol{\beta} \mid \mathbf{w})}{\partial \boldsymbol{\beta}_i \partial \mathbf{w}_j}\right],$$

avaliados em $\hat{\boldsymbol{\beta}}$ e $\mathbf{w}_0$.

Nessa situação, as matrizes $\mathbf{Q}$ e $\boldsymbol{\Delta}$ são dadas por

$$-\mathbf{Q} = \frac{(\mathbf{X}'\mathbf{X} + k\mathbf{I})}{\sigma^2}$$

$$\mathbf{\Delta} = \frac{\mathbf{X}'\mathbf{D}(\mathbf{e}^*)}{\sigma^2}$$

onde $\mathbf{e}^*$ é o vetor de resíduos em cristas, isto é, $\mathbf{e}^* = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*$, $\hat{\boldsymbol{\beta}}^*$ é o estimador em cristas e $\mathbf{D}(\mathbf{e}^*) = diag(\mathbf{e}_1^*, ..., \mathbf{e}_n^*)$. A curvatura é obtida como:

$$\begin{aligned} C_d &= 2|\mathbf{d}'\mathbf{F}\mathbf{d}| \\ &= 2|\mathbf{d}'\mathbf{\Delta}'\mathbf{Q}^{-1}\mathbf{\Delta}\mathbf{d}| \\ &= \frac{2|\mathbf{d}'\mathbf{D}(\mathbf{e}^*)\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{D}(\mathbf{e}^*)\mathbf{d}|}{\sigma^2}. \end{aligned}$$

A curvatura máxima será obtida a partir do maior autovalor $\lambda_{\max}^*$ de

$$\mathbf{D}(\mathbf{e}^*)\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{D}(\mathbf{e}^*),$$

e é dada por

$$C_{\max} = \frac{2\lambda_{\max}^*}{\sigma^2}. \qquad (4.3)$$

De maneira análoga, para o modelo (4.1) com σ^2 conhecido e suposição de homogeneidade na variância dos erros relaxada via pequenos distúrbios, será feita uma análise de influência no método de estimação em cristas através da medida alternativa $l_{\max}$. A função $LD_{\Re}^*$, que representa a versão em cristas de LD^* pode ser escrita como:

$$LD_{\Re}^* = -2\left[L_p(\hat{\boldsymbol{\beta}}^*) - \mathcal{L}_p(\hat{\boldsymbol{\beta}}_w^* \mid \mathbf{w})\right],$$

onde $L_p(\hat{\boldsymbol{\beta}}^*)$ e $\mathcal{L}_p(\hat{\boldsymbol{\beta}}_w^* \mid \mathbf{w})$ são respectivamente as funções de verossimilhança sem e com distúrbios.

Admitindo, por exemplo, que somente o primeiro elemento amostral sofreu distúrbio de modo que $\mathrm{w}_1^* = 1 + \mathrm{w}_1$ e $\mathrm{w}_i^* = 1$ para $i > 1$, teremos

$$\frac{\partial LD^*_{\Re}}{\partial w_1} = 1 - \frac{(e^*_1)^2}{\sigma^2},$$

onde e^*_1 representa o primeiro elemento do vetor de resíduos em cristas. Se introduzirmos distúrbios nos n elementos, obteremos:

$$\frac{\partial LD^*_{\Re}}{\partial w_i} = 1 - \frac{(e^*_i)^2}{\sigma^2}, \text{ para } i = 1, 2, ..., n,$$

onde e^*_i é o *i-ésimo* elemento do vetor de resíduos em cristas.

Assim, a máxima inclinação será dada por

$$\left|\nabla LD^*_{\Re}\right| = \left[\sum_{i=1}^{n}\left(1 - \frac{(e^*_i)^2}{\sigma^2}\right)^2\right]^{1/2}. \qquad (4.4)$$

Com relação a uma interpretação adequada dessas medidas, dúvidas podem surgir sobre a partir de quais valores essas medidas poderiam ser consideradas grandes o suficiente de modo a sugerir uma investigação futura. Cook (1986) sugere que $C_{\max} = 2$ pode ser usado como um valor limite. Contudo, Billor e Loynes (1993) apontaram o valor $\sqrt{2n + 4(14n)^{1/2}}$ como relevante na determinação de influência local. Este ponto de corte foi obtido por Billor e Loynes (1993), utilizando uma aproximação para a média e variância de $A = \sum_{i=1}^{n}\left(1 - \frac{(e^*_i)^2}{\sigma^2}\right)^2$. Verificou-se que $E(A) = 2n$, $Var(A) = 56n$, de modo que $E(A) + 2DP(A) = 2n + 4\sqrt{14n}$. Por outro lado, Loynes (1997) sugeriria que $2p$ é um valor mais apropriado do que aqueles citados anteriormente (2 e $\sqrt{2n + 4(14n)^{1/2}}$).

4.5 – Exemplo

Billor e Loynes (1999) analisaram o banco de dados "Tobacco data" (Billor, 1992) com a finalidade de aplicar tanto a versão adaptada da medida LD de Cook, quanto a versão alternativa desenvolvida pelos mesmos, todas no contexto da regressão em cristas. Para uma amostra de trinta marcas de tabaco, um modelo de regressão linear múltipla com variável resposta $\mathbf{y}$, quantidade de calor produzida pelo tabaco durante o processo de fumo, em função das porcentagens de concentração de quatro componentes importantes $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ e $\mathbf{X}_4$, foi adotado.

Os autores consideraram a variância dos erros como iguais, ou seja, mantiveram a suposição da homogeneidade de variâncias, $\text{var}(\varepsilon_i) = \sigma^2$, $i = 1, 2, ..., 30$. Para esse conjunto de dados, o número condicional $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$ é igual a 22.293 e a regressão em cristas foi utilizada para reduzir o efeito da multicolinearidade na estimação de $\boldsymbol{\beta}$.

A estimativa obtida foi $\hat{\boldsymbol{\beta}}^* = (760{,}5247;\ 1163{,}277;\ -441{,}311;\ -475{,}754;\ 867{,}482)'$ utilizando, para k, o valor proposto por Hoerl-Kennard, $k = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}}$, onde $\hat{\sigma}^2$ e $\hat{\boldsymbol{\beta}}$ são os estimadores de máxima verossimilhança de σ^2 e $\boldsymbol{\beta}$, respectivamente. No cálculo da medida de influência de Cook, a curvatura máxima, que corresponde ao maior autovalor $\lambda^*_{\max}$ de $\mathbf{D}(\mathbf{e}^*)\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{D}(\mathbf{e}^*)$, foi encontrada a partir da equação (4.3), utilizando $\hat{\sigma}^2$, estimativa de máxima verossimilhança de σ^2. Nessas condições, a curvatura máxima obtida foi $C_{\max} = 2(4113{,}84)/2290{,}30 = 3{,}59$. Com base nesse valor, conclui-se pela existência de uma moderada sensibilidade local nos dados, de acordo com o critério de Cook $(C_{\max} > 2)$. Contudo, o critério de Loynes (1997) indica

$(C_{\max} = 3{,}59 < 8)$, pelo fato de $p = 4$. Assim, a partir dos critérios de Loynes e Cook, podemos dizer que existe uma sensibilidade local moderada na base de dados.

Para o cálculo da inclinação máxima, os resíduos em cristas foram obtidos. A inclinação máxima dada pela equação (4.4) não sugere qualquer influência local, pelo fato de $l_{\max} = 9{,}2 < 11{,}9$, calculado como $\sqrt{2n + 4(14n)^{1/2}}$. Contudo, uma avaliação individual das componentes $l_i = 1 - \frac{(e_i^*)^2}{\hat{\sigma}^2}$ de $l_{\max}$ fornece alguma idéia de quais casos são mais afetados pelos seus respectivos distúrbios. Os valores individuais estão organizados na Tabela 4.1. Nela, podemos perceber que os casos 4, 8 e 14 possuem valores de l_i bem maiores que os demais. Essas observações são as que mais contribuem para a inclinação máxima e são visivelmente afetadas pela introdução do distúrbio na variância.

Tabela 4.1 - Os valores individuais l_i para os dados de tabaco

Caso (i)	l_i	Caso (i)	l_i	Caso (i)	l_i
1	0,6959	11	0,2227	21	0,9999
2	0,9445	12	0,5636	22	0,9853
3	0,9876	13	0,9863	23	0,9893
4	7,5035	14	2,5753	24	0,9963
5	0,5397	15	0,6294	25	0,9809
6	0,0162	16	0,8876	26	0,9611
7	0,7106	17	0,2401	27	0,9989
8	2,1131	18	0,8277	28	0,9350
9	0,8159	19	0,8714	29	0,9833
10	0,8288	20	0,0268	30	0,1969

Fonte: Billor, N. e Loynes, R. M. (1999).

Billor e Loynes (1999) identificaram, também, que os casos 4 e 8 exibiram resíduos externamente studentizados elevados: $r_4 = 3{,}738$ e $r_5 = 2{,}448$. Assim, a partir dos valores individuais l_i, acredita-se que esses casos tenham sido afetados pela introdução de distúrbios, gerando dúvidas sobre a homogeneidade da variância dos erros.

Capítulo 5

Regressão Robusta em Cristas

5.1 - Introdução

Nesse capítulo será apresentado um método alternativo de estimação, qual seja, o método robusto de estimação em cristas introduzido em Lawrence e Marsh (1984).

Nesse trabalho, os autores comparam diferentes combinações dos métodos robustos e regressão em cristas com o objetivo de estimar os parâmetros de um modelo para o número de fatalidades na indústria mineradora de carvão dos Estados Unidos em função de seis variáveis explicativas.

De acordo com os autores, essa combinação entre as técnicas de regressão em cristas e robusta é necessária pelo fato da ocorrência simultânea de multicolinearidade e outliers nos dados.

O modelo utilizado é expresso na forma log-linear e é dado por

$$\ln Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + ... + \beta_6 \ln X_{6i} + e_i,$$

com

Y = número de fatalidades de explosões de gás e poeira;

X_1 = número de máquinas de corte em uso;

X_2 = produção de carvão em toneladas por horas de trabalho;

X_3 = número médio de mineiros trabalhando no subsolo;

X_4 = número de máquinas móveis de carga em uso;

X_5 = número de máquinas de mineração contínuas em uso;

X_6 = total anual de horas de trabalho da produção.

Como o número de máquinas de corte está positivamente associado às técnicas de perfuração e abertura com explosivos, espera-se que o coeficiente de regressão da variável X_1 tenha valor positivo. A produção de carvão em toneladas por horas de trabalho mede a intensidade dessa atividade. Assim, também é esperado que o coeficiente da variável X_2 seja positivo. Quanto maior o número médio de mineiros trabalhando no subsolo, maior é o número esperado de fatalidades. Logo, espera-se que o coeficiente da variável X_3 seja positivo. A partir de 1950, o emprego crescente tanto de máquinas móveis de carga X_4, quanto de máquinas de mineração contínuas X_5 foi responsável pela modernização e mecanização dessa atividade. Dessa maneira, os valores dos coeficientes de regressão associados a X_4 e X_5 deveriam ser negativos.

Finalmente, é esperado um valor positivo para o coeficiente da variável X_6, pois quanto maior o total anual de horas de trabalho, maior é o número esperado de fatalidades.

Os dados de fatalidades, Y, são provenientes do Departamento de Trabalho de Minas e Segurança dos Estados Unidos. Os valores das variáveis explicativas foram obtidos da Agência de Minas dos Estados Unidos e do Departamento de Energia dos Estados Unidos. A base de dados consiste de 64 observações anuais, de 1915 a 1978.

Segundo os autores, foi identificada uma forte correlação entre as variáveis explicativas, maior que a correlação com a variável dependente. Essa evidência, e outras que serão ainda apresentadas, sugere um alto grau de multicolinearidade nos dados.

Procedeu-se, inicialmente, ao ajuste do modelo através do procedimento de mínimos quadrados com base nos dados não padronizados, obtendo-se

$$\begin{aligned}\ln \hat{Y}_i = &-58{,}45 + 3{,}37 \ln X_{1i} + 4{,}35 \ln X_{2i} - 1{,}48 \ln X_{3i} - 0{,}003 \ln X_{4i} \\ &- 0{,}166 \ln X_{5i} + 2{,}30 \ln X_{6i}.\end{aligned} \qquad (5.1)$$

Na Tabela 5.1 os coeficientes padronizados correspondentes estão no grupo SEM PESO, na coluna MQ.

O coeficiente de explicação R^2 referente a esse ajuste é de 0,5339 com valor da estatística F de 10,881 (altamente significante), com p-valor menor que 0,0001. Contudo, apenas um dos coeficientes estimados apresentou estatística t significante para qualquer nível de significância razoável. Além disso, era esperado que a variável X_3 apresentasse um coeficiente positivo, fato que não ocorreu. Finalmente, os autores perceberam que quatro das seis variáveis tiveram FIV maior que 5 sugerindo, dessa forma, a existência de multicolinearidade.

Um outro problema de interesse real diz respeito à existência de outliers nos dados. Para detectar a influência de outliers nos dados, Lawrence e Marsh (1984) utilizaram a distância de Cook. Dessas 64 observações, seis apresentaram distância de

Cook maior ou igual a 0,140 sugerindo, assim, a presença de outliers. Por esse motivo, vários métodos robustos combinados com técnicas de regressão em cristas foram utilizados como forma de lidar com situações nas quais outliers e multicolinearidade ocorrem simultaneamente.

5.2 – Regressão Robusta em Cristas

Com o objetivo de minimizar o efeito da multicolinearidade e de pontos discrepantes, foi utilizado um procedimento robusto de regressão sobre cristas, que realiza ponderações nas variáveis explicativas. A idéia de combinar métodos de regressão robusta com regressão em cristas tem sido discutida por inúmeros pesquisadores.

Uma questão de interesse, e ainda sem consenso, é se o parâmetro k da regressão em cristas deveria ou não receber um peso. Lawrence e Marsh (1984) não atribuíram peso algum ao parâmetro k mas mantiveram a matriz diagonal de pesos, $\mathbf{W}$, e a matriz diagonal dos parâmetros viesados, denotada por K, separados segundo o modelo de regressão em cristas robusta:

$$\hat{\boldsymbol{\beta}}_K = (\mathbf{X}'\mathbf{W}\mathbf{X} + K)^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}\,. \qquad (5.2)$$

Esse modelo equivale a multiplicar cada valor observado das variáveis explicativas e resposta pela raiz quadrada dos respectivos pesos.

Para determinar os pesos w_i que constituem os elementos diagonais da matriz $\mathbf{W}$, quatro métodos alternativos de distribuição de pesos foram adotados: método de Huber, método de Andrews, método de Hampel e método de Ramsay.

Tais procedimentos são descritos em Montgomery e Peck (1982, pág. 367). De modo geral, uma classe de estimadores robustos é formada por estimadores que minimizam uma particular função ρ dos resíduos, ou seja, obtem-se $\hat{\boldsymbol{\beta}}$ que minimiza

$$\sum_{i=1}^{n} \rho\left[\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)/s\right], \tag{5.3}$$

em que s é um estimador robusto do parâmetro de escala σ^2.

Os autores utilizaram o valor de s usual dado por

$$s = mediana\ \frac{\left|e_i - mediana\left(e_i\right)\right|}{0{,}6745},$$

em que e_i é o resíduo do modelo ajustado através do procedimento de regressão L_1, que minimiza a soma dos valores absolutos dos resíduos.

A minimização da expressão (5.3) resulta em um sistema de $k+1$ equações,

$$\sum_{i=1}^{n} x_{ij}\, \psi\left[\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{s}\right] = 0, \tag{5.4}$$

$j = 0,1,...,k$, sendo ψ a derivada da função ρ, x_{ij} o *i-ésimo* valor da *j-ésima* variável explicativa e $x_{i0} = 1$.

De modo geral, a função ψ é não linear e o sistema (5.4) deve ser resolvido por métodos iterativos. Montgomery e Peck (1982) sugerem o uso do procedimento iterativo de mínimos quadrados ponderados, que fornece como solução o estimador

$$\hat{\boldsymbol{\beta}}_1 = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{y},$$

em que $\mathbf{W}$ é a matriz diagonal de pesos, $n \times n$, cujos elementos da diagonal principal são:

$$w_i = \begin{cases} \dfrac{\psi\left[\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)/s\right]}{\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)/s} & se\ y_i \neq \mathbf{x}_i'\hat{\boldsymbol{\beta}}_0 \\ 1 & se\ y_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_0, \end{cases}$$

para $\hat{\boldsymbol{\beta}}_0$ estimativa inicial de $\boldsymbol{\beta}$.

Nos passos seguintes, os pesos são recalculados substituindo-se $\hat{\boldsymbol{\beta}}_0$ por $\hat{\boldsymbol{\beta}}_1$ e o procedimento é repetido sucessivamente até que a convergência seja atingida.

Diferentes procedimentos de regressão robusta são obtidos para diferentes formas de função ψ. Tal função controla o peso que é dado a cada resíduo no cálculo do estimador de $\boldsymbol{\beta}$.

Para os procedimentos utilizados por Lawrence e Marsh (1984), temos:

- **Critério de Hubber com coeficiente t:**

$$\rho(z) = |z|t - \frac{1}{2}t^2,$$

$$\psi(z) = t \,\text{sinal}(z),$$

$$w(z) = \frac{t}{|z|}, \quad para\ |z| > t.$$

- **Critério de Ramsay com coeficiente a:**

$$\rho(z) = a^{-2}\left[1 - \exp(-a|z|)\cdot(1 + a|z|)\right],$$

$$\psi(z) = z\exp(-a|z|),$$

$$w(z) = \exp(-a|z|), \quad para\ |z| < \infty.$$

- **Critério de Andrew com coeficiente a:**

$$\rho(z) = a\left[1 - \cos\left(\frac{z}{a}\right)\right],$$

$$\psi(z) = sen\left(\frac{z}{a}\right),$$

$$w(z)=\frac{sen(z/a)}{z/a}, \quad para\ |z|\le a\pi .$$

- **Critério de Hampel com coeficientes a, b e c:**

$$\rho(z)=\frac{a\left(c|z|-\frac{1}{2}z^2\right)}{c-b}-\frac{7}{6}a^2,$$

$$\psi(z)=\frac{a\ \text{sinal}(z)\cdot(c-|z|)}{c-b},$$

$$w(z)=\frac{a(c-|z|)}{|z|(c-b)}, \quad para\ b\le|z|\le c .$$

Montgomery e Peck (1982, pág. 369) sugerem um coeficiente Huber de $t=2{,}0$, enquanto Vinod e Ullah (1981), citados por Lawrence e Marsh (1984), sugerem que o mesmo seja 1,345. Na análise realizada, os autores utilizaram $t=1{,}5$.

Para o parâmetro do método de Andrew, c, o valor adotado por Lawrence e Marsh (1984) para a regressão foi de 1,339, enquanto que Hogg (1979), citado pelos mesmos, sugere um valor de 1,5 ou 2,0. Vale ressaltar que se o parâmetro de escala é conhecido, o valor de 1,339 exige um acréscimo de 5%.

Para o parâmetro de Hampel, foram utilizados os valores padrão: $a=1{,}7$, $b=3{,}4$ e $c=8{,}5$.

Com relação ao coeficiente de Ramsay, Ramsay (1977) analisou os valores 0,1, 0,3, 0,5 e 1,0 para o parâmetro a, concluindo que 0,3 gerou resultados melhores.

Após cada observação ter recebido um peso por meio da matriz **W**, segundo algum método de regressão robusta, os dados foram padronizados de modo que as matrizes **X'WX** e **X'WY** se transformaram em matrizes de correlação.

Na determinação do valor de k para a regressão em cristas, quatro métodos foram utilizados: o estimador de Lindley e Smith (LS), o estimador de Hoerl, Kennard e Baldwin (HKB), o estimador de Lawless e Wang (LW) e o estimador generalizado (GEN).

Lindley e Smith (1972) basicamente propuseram um estimador bayesiano para k, obtido a partir dos dados, após o ajuste do modelo de regressão em que $\beta_j \sim (\xi, \sigma_\beta^2)$. Neste caso, o parâmetro k é da forma $\frac{\sigma^2}{\sigma_\beta^2}$, onde σ_β^2 é a variância da distribuição a priori de β_j e k é estimado por $k^* = \frac{s^2}{s_\beta^2}$, conforme expressão (38) de Lindley e Smith (1972). Os autores apresentaram uma condição para a qual o erro quadrático médio do estimador proposto (LS) é menor que o de mínimos quadrados. Concluem, ainda, que a chance de que a condição seja satisfeita tende rapidamente a um quando n cresce.

Hoerl, Kennard e Baldwin (1975) propuseram o estimador HKB, $k = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}}$, no qual $\hat{\sigma}^2$ e $\hat{\boldsymbol{\beta}}$ são estimados por mínimos quadrados e demonstraram também, por meio de simulações, que esse estimador produziu erro quadrático médio para os coeficientes de regressão menor quando comparado com o de mínimos quadrados.

Lawless e Wang (1976) sugerem um estimador ligeiramente diferenciado do estimador de Hoerl, Kennard e Baldwin. Escrevendo o estimador HKB como $k = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p} \hat{\alpha}_i^2}$, para $\hat{\alpha}_i = \hat{\beta}_i$, os autores propõem o uso do estimador $k = \frac{p\hat{\sigma}_2}{\sum_{i=1}^{p} \lambda_i \hat{\alpha}_i^2}$, no qual os valores λ_i são os autovalores da matriz $\mathbf{X}'\mathbf{X}$. Após simulações os autores

concluíram por um bom desempenho desse estimador, no que diz respeito ao critério do erro quadrático médio.

O estimador generalizado GEN foi proposto por Hoerl e Kennard (1970) como uma forma de extensão ao procedimento de regressão em cristas que permite separar os parâmetros viesados para cada variável explicativa. Basicamente o modelo linear $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ é reescrito em sua forma canônica $\mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$, por meio de algumas transformações. Em termos de sua forma canônica, o estimador em cristas generalizado é solução de $(\mathbf{Z}'\mathbf{Z} + K)\hat{\boldsymbol{\alpha}}_{GR} = \mathbf{Z}'\mathbf{y}$. E em termos de seu modelo original, os coeficientes em cristas generalizados são dados por $\hat{\boldsymbol{\beta}}_{GR} = \mathbf{T}\hat{\boldsymbol{\alpha}}_{GR}$.

As estimativas obtidas encontram-se organizadas na Tabela 5.1.

Tabela 5.1 – Estimativas dos Coeficientes de Regressão						
		MQ	LS	HKB	LW	GEN
SEM PESO	X1	0,765	0,666	0,668	0,573	0,698
	X2	0,565	0,345	0,348	0,222	0,431
	X3	-0,248	-0,117	-0,119	-0,049	-0,163
	X4	-0,005	-0,003	-0,003	0,001	-0,049
	X5	-0,377	-0,299	-0,300	-0,239	-0,357
	X6	0,375	0,185	0,187	0,131	0,196
		MQ	LS	HKB	LW	GEN
HUBER	X1	0,491	0,290	0,292	0,317	0,357
	X2	0,252	0,140	0,141	0,162	0,178
	X3	0,120	0,174	0,173	0,156	0,143
	X4	-0,072	-0,060	-0,060	-0,062	-0,071
	X5	-0,465	-0,370	-0,372	-0,392	-0,365
	X6	-0,038	0,172	0,170	0,151	0,141
		MQ	LS	HKB	LW	GEN
ANDREWS	X1	1,015	0,613	0,618	0,332	0,882
	X2	0,429	0,246	0,248	0,112	0,270
	X3	0,372	-0,080	-0,083	0,068	-0,159
	X4	-0,110	-0,096	-0,096	-0,086	-0,123
	X5	-0,720	-0,647	-0,648	-0,560	-0,612
	X6	-1,189	-0,202	-0,207	0,033	-0,394
		MQ	LS	HKB	LW	GEN
HAMPEL	X1	0,781	0,462	0,466	0,278	0,629
	X2	0,439	0,250	0,252	0,122	0,270
	X3	0,487	-0,024	-0,026	0,083	-0,048
	X4	-0,112	-0,098	-0,098	-0,089	-0,128
	X5	-0,741	-0,662	-0,663	-0,572	-0,630
	X6	-1,075	-0,114	-0,118	0,057	-0,262
		MQ	LS	HKB	LW	GEN
RAMSAY	X1	0,866	0,525	0,530	0,403	0,756
	X2	0,339	0,195	0,196	0,143	0,222
	X3	0,244	0,027	0,025	0,089	-0,067
	X4	-0,097	-0,084	-0,084	-0,079	-0,112
	X5	-0,569	-0,498	-0,499	-0,460	-0,467
	X6	-0,695	-0,046	-0,049	0,051	-0,178

Fonte: Lawrence e Marsh (1984).

5.3 – Resultados

Marquardt (1970) estabelece que os fatores de inflação de variância (FIV) para o procedimento de regressão em cristas são os elementos da diagonal principal de

$$(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\cdot(\mathbf{X}'\mathbf{X})\cdot(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1},$$

em que $\mathbf{X}'\mathbf{X}$ está escrito na forma de matriz de correlação e k é o valor utilizado na expressão do estimador da regressão em cristas.

As conclusões apresentadas por Lawrence e Marsh (1984) foram baseadas no fator de inflação da variância (FIV) e nas estimativas dos coeficientes da regressão. Inicialmente, todos os métodos de regressão em cristas obtidos através da utilização das diferentes expressões de k, reduziram os FIV's como esperado. Os métodos LW e GEN foram os que forneceram uma maior redução dos FIV's. Os métodos LS e HKB apresentaram resultados bem parecidos tanto em termos de FIV's, quanto em termos de coeficientes estimados.

Apesar de os FIV's terem sido reduzidos nos métodos L-W e GEN, no caso particular do método L-W, os FIV's ficaram muito próximos uns dos outros, enquanto que no método GEN, os FIV's gerados variaram consideravelmente.

Em termos de credibilidade das estimativas para o modelo de fatalidades, o método LW foi o único que consistentemente produziu os sinais esperados para as estimativas dos coeficientes do modelo de regressão robusto. Nenhum método produziu resultados razoáveis nos casos em que não foram utilizados pesos.

De certa forma, a técnica robusta Huber parece ser a que apresenta resultados consistentes com aqueles que eram esperados. Contudo, o método LW que produziu consistentemente os sinais corretos sob todas as técnicas robustas, apresentou FIV maior que cinco sob a técnica Huber. Isso sugere que alguma multicolinearidade ainda persistiu após a aplicação desse método. Esse parece não ser o caso para o estimador LW sob as técnicas robustas Andrews ou Hampel. Contudo, essas técnicas robustas não proporcionaram resultados razoáveis quando outros estimadores em cristas foram

usados. Dessa forma, a técnica robusta Huber parece ser a mais estável e a mais confiável para esses dados.

Consequentemente, a técnica robusta Huber usada em conjunto com o estimador LW forneceu o melhor o melhor método de ajuste no modelo que relaciona o número de fatalidades na indústria mineradora de carvão dos Estados Unidos em função das variáveis explicativas descritas.

Capítulo 6

Aplicações

6.1 – Introdução:

O presente capítulo tem por objetivo a aplicação de algumas das técnicas de análise apresentadas anteriormente ao conjunto de dados reais do projeto: **Relação Estrutura-Atividade de Anestésicos Locais N,N [Dimetilamina] Etil Benzoatos Para-Substituídos**, (RAE-CEA-9710, André, Elian e Bruscato, 1997).

O projeto é originalmente da área farmacológica e investiga o efeito de diversos tipos de anestésicos locais sobre o coração de ratos. O interesse desse estudo consistiu em verificar quais características físico-químicas da molécula de determinada droga influenciam mais em sua potência tóxica, definida como a dose de droga necessária para

ocorrer uma redução de 30% na freqüência do átrio. Para tal, foram utilizados setenta e dois ratos, homogêneos entre si, divididos em quatorze grupos, contendo de três a oito ratos. Cada grupo foi submetido a uma droga diferente e a potência tóxica calculada após a realização de um experimento, descrito no referido trabalho.

6.2 – Descrição das Variáveis:

Foram consideradas dez variáveis independentes, constituídas exatamente por dez características físico-químicas das próprias drogas, de modo que seus valores são determinados sem erro. Estão divididas em quatro grupos, de acordo com o efeito ao qual se referem.

As variáveis do grupo do *efeito estérico* mediam um comprimento relacionado ao grupamento substituinte. São elas:

- **L**: comprimento do grupo substituinte ao longo do eixo da ligação com o esqueleto da molécula (medido em Ângstron);
- **B1** e **B4**: larguras do comprimento substituinte a partir do eixo da ligação, perpendiculares a ele (medidas em Ângstron).

As variáveis do grupo do *efeito eletrônico* mediam o efeito de dispersão de carga em torno do anel benzênico e eram definidas como:

- **F**: componente de campo (adimensional);
- **R**: componente de ressonância (adimensional);
- **SIGMA**: constante de Hammet – combinação linear das duas anteriores (adimensional);

- **C (CARBONILA)**: freqüência de estiramento da carbonila medida por espectroscopia no infra-vermelho (em cm^{-1}).

As variáveis do grupo do *efeito hidrofóbico* mediam quanto as moléculas da droga misturam-se na água e eram definidas como:

- **LOG.PAPP**: logaritmo do coeficiente de partição óleo-água medido (adimensional);
- **PI**: coeficiente de partição óleo-água calculado (adimensional).

A variável do grupo de *outros efeitos* é:

- **MR4**: refratividade molar (adimensional),

e a variável resposta é dada por:

- **POTÊNCIA**: $-\log(DE_{30})$, onde DE_{30} é a dose de droga necessária para ocorrer uma redução de 30% na freqüência do átrio em relação ao controle (adimensional).

Na análise da relação entre a variável resposta e as variáveis independentes foi utilizado um modelo de regressão linear múltipla. Desconfiava-se, no entanto, da presença de multicolinearidade e, por esse motivo, André, C. D. S. de, Elian, S. N., Bruscato, A. (1997) consideraram três grupos de variáveis independentes, obtidos com base na matriz de correlações (Tabela A.1, Apêndice A).

- Grupo 1: formado por todas as variáveis independentes consideradas no projeto;
- Grupo 2: formado pelas variáveis independentes mais correlacionadas linearmente (coeficientes de correlação linear de Pearson) com a variável

potência em cada um dos quatro grupos de efeitos. São elas: B4, SIGMA, LOG.PAPP e MR4;

- Grupo 3: Foi construído um dendograma (Tabela A.2, Apêndice A) de modo a agrupar as variáveis independentes altamente correlacionadas entre si. Realizado um corte ao nível 7,5, foram obtidos cinco agrupamentos: LOG.PAPP e PI; L, MR4 e B4; R e C; B1 e SIGMA e F. O grupo 3 foi formado pelas variáveis independentes mais correlacionadas linearmente com a variável potência em cada um dos agrupamentos obtidos. Essas variáveis são B4, SIGMA, F, R e LOG.PAPP.

Como os grupos 2 e 3 contém variáveis independentes altamente correlacionadas entre si, aplicaremos o procedimento de regressão em cristas e algumas das técnicas de diagnóstico apresentadas nos capítulos anteriores. Para tal, as análises serão desenvolvidas no pacote computacional R. Alguns comandos utilizados encontram-se no Apêndice B.

6.3 - Análises

Inicialmente, aplicaremos o procedimento de regressão em cristas calculando algumas das medidas de diagnóstico para o grupo 2. Posteriormente, o mesmo será feito para o grupo 3.

- **Grupo 2:**

Os coeficientes do modelo de regressão ajustado através do procedimento de mínimos quadrados para o grupo 2 encontram-se na Tabela 6.1.

Tabela 6.1 - Ajuste do Modelo por Mínimos Quadrados - Grupo 2

	Estimativas	Erro Padrão	Valor da estatística t	p-valor
(Intercepto)	3,11	0,11	27,83	<2e-16
B4	0,11	0,08	1,31	0,1932
SIGMA	-0,60	0,09	-6,85	2,81e-09
LOG.PAPP	0,31	0,04	8,45	3,78e-12
MR4	-0,29	0,13	-2,19	0,0318

Na análise realizada, é possível perceber que o coeficiente da variável B4 não é estatisticamente significante ao nível de 0,05. Uma possível razão para esse fato seria a existência de multicolinearidade nos dados. A fim de detectá-la, foram calculados os FIV's associados a cada variável, que se encontram na Tabela 6.2.

Tabela 6.2 - Fator de Inflação da Variância - Procedimento de Regressão de MQ - Grupo 2

Variável	B4	SIGMA	LOG.PAPP	MR4
FIV	28,75	1,83	3,62	26,17

Assim, podemos perceber que as variáveis B4 e MR4 são altamente correlacionadas com uma ou mais variáveis independentes por apresentarem FIV maior que 5.

Uma outra forma de avaliar a multicolinearidade é por meio do número condicional κ, que é obtido pela razão $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$, onde λ_{max} é o maior autovalor da matriz $(\mathbf{X'X})$, na sua forma de correlação, enquanto que λ_{min} reflete o menor autovalor da matriz $(\mathbf{X'X})$. Seus autovalores são dados por: 3,1562; 0,6171; 0,2080; 0,01865.

Logo, $\kappa = \frac{\lambda_{max}}{\lambda_{min}} = \frac{3,1562}{0,01865} = 169,2$. Como $100 < \kappa < 1000$, podemos concluir pela existência de sensível multicolinearidade nos dados.

Por esse motivo será adotado o procedimento de regressão em cristas, determinando, inicialmente, o valor para k através da análise do traço, para k variando de zero a dois. O traço corresponde ao gráfico dos valores das estimativas em função de k.

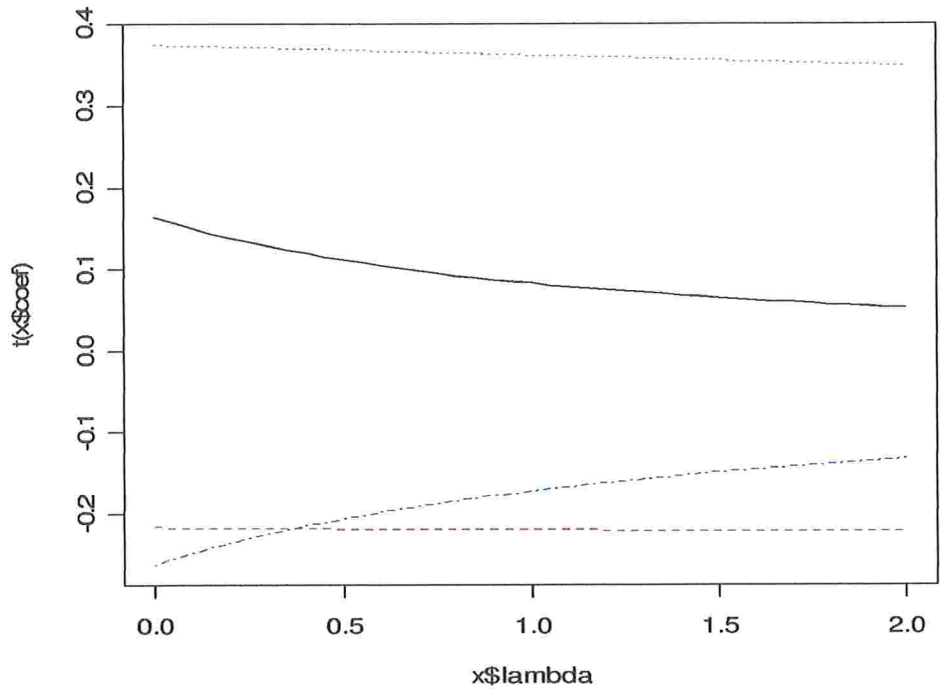


Figura 6.1 – Traço para as variáveis: B4, SIGMA, LOG.PAPP e MR4 - Grupo 2

Através da Figura 6.1, podemos perceber que a partir de $k = 1{,}5$ os coeficientes de regressão tendem a se estabilizar. Assim, esse valor será escolhido e um modelo de regressão em cristas para os dados do grupo 2 pode ser expresso por:

$$\hat{Y} = 3{,}19 + 0{,}04 \cdot \text{B4} - 0{,}61 \cdot \text{SIGMA} + 0{,}29 \cdot \text{LOG.PAPP} - 0{,}16 \cdot \text{MR4}.$$

Os valores dos FIV's para as variáveis do grupo 2, após ajuste de um modelo de regressão em cristas, decresceram consideravelmente e são dados por:

Tabela 6.3 - Fator de Inflação da Variância - Procedimento de Regressão em Cristas - Grupo 2

Variável	B4	SIGMA	LOG.PAPP	MR4
FIV	0,09	0,02	0,04	0,09

Foram calculados os elementos da diagonal principal da matriz $\mathbf{H}^* = \mathbf{Z}(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^*)^{-1}\mathbf{Z}'$, obtidos a partir da expressão da seção 5.3, com função similar à matriz "hat" na estimação por mínimos quadrados. As observações 64, 65 e 66 foram consideradas como as mais influentes, com valor $h_i^* = 0{,}169$. Isso pode ser verificado na Figura 6.2. Ao analisarmos suas características, percebemos que esses ratos apresentaram os maiores valores de SIGMA (0,72) e foram os únicos três ratos com valores negativos em LOG.PAPP (-0,70).

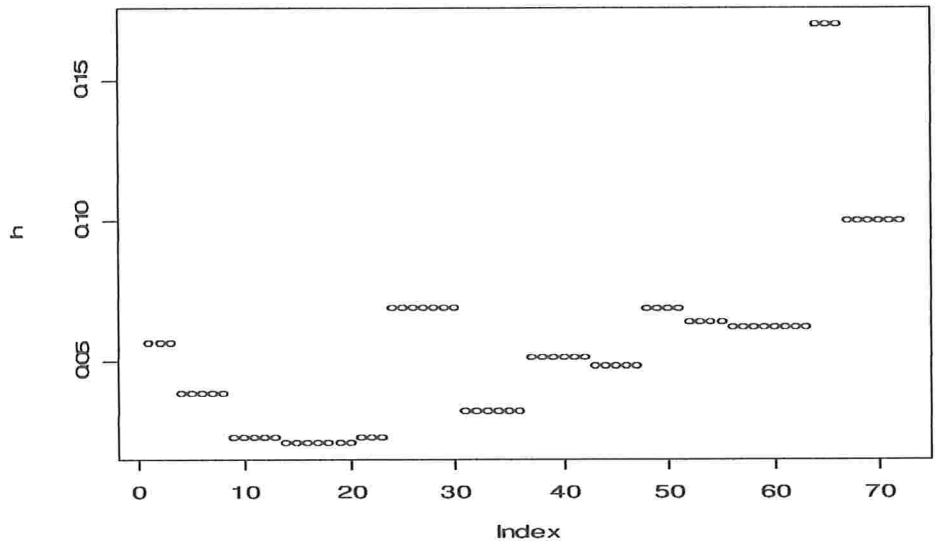


Figura 6.2 - Valores da diagonal principal da matriz H* - Grupo 2

É necessário enfatizar, ainda, que, por meio desse modelo de regressão em cristas, não foram detectados pontos influentes através da medida D_i^*, no que diz respeito ao grupo 2. Realizado o ajuste pelo método dos mínimos quadrados, de acordo com as medidas tradicionais, oito seriam as observações influentes: 21, 44, 64, 65, 68, 69, 70 e 71.

Com relação às medidas de influência local, $\lambda_{max}^* = 0{,}046842$ e a curvatura máxima foi obtida resultando em $C_{max} = \frac{2 \cdot \lambda_{max}^*}{\hat{\sigma}^2} = \frac{2 \cdot 0{,}046842}{0{,}03910028} = 2{,}40$. Dessa forma, alguma sensibilidade local existe nos dados, de acordo com o critério de Cook $(C_{max} > 2)$. Contudo, pelo critério sugerido por Loynes (1997), $C_{max} = 2{,}40 < 8$, pelo fato de $p = 4$. Assim, a partir dos critérios de Cook e Loynes, podemos concluir pela existência de uma sensibilidade moderada.

Após a obtenção do maior autovalor λ_{max}^*, seu autovetor associado também nos fornece informações sobre os pontos mais influentes. Neste caso, as coordenadas com maiores valores correspondem aos pontos mais influentes. Para o grupo 2, destacamos os ratos de números: 44, 46, 21, 67, 45 e 47. Todos apresentaram valores superiores a $|0{,}2|$, fato que pode ser verificado na Figura 6.3.

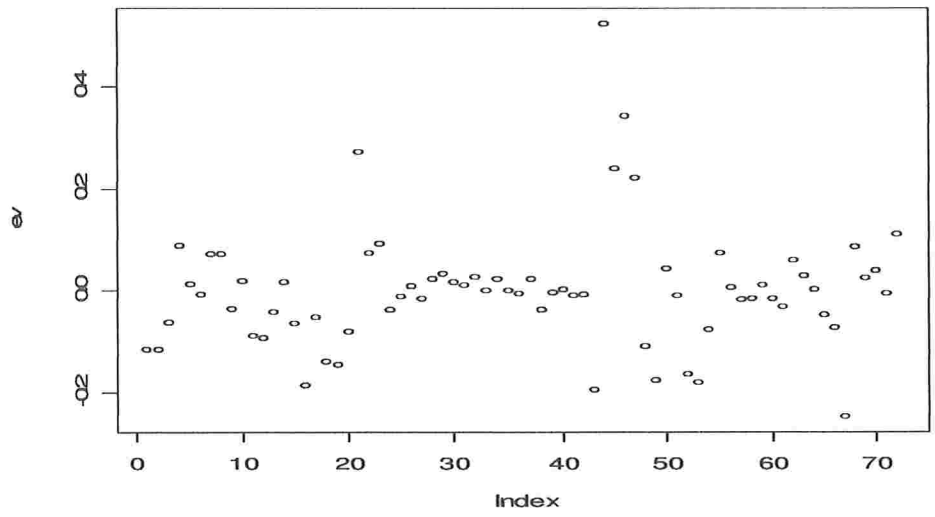


Figura 6.3 - Análise da Influência pelos componentes do autovetor associado a λ^*_{max} - Grupo 2

A inclinação máxima l_{max} foi determinada, após obtenção dos resíduos do modelo de regressão em cristas, de acordo com

$$\left|\nabla LD^*_{\Re}\right| = \left[\sum_{i=1}^{n}\left(1-\frac{\left(\mathbf{e}_i^*\right)^2}{\hat{\sigma}^2}\right)^2\right]^{1/2},$$

obtendo-se

$$\left|\nabla LD^*_{\Re}\right| = 12{,}72\,.$$

Um critério de avaliação para essa medida de inclinação máxima é dado pela raiz quadrada de $2n+4\sqrt{14n}$ que, para os dados do grupo 2, equivale a 16,46.

Assim, como $l_{max} = 12{,}72 < 16{,}46$, temos que l_{max} não sugere sensibilidade local para os dados do grupo 2. Mas, ainda sim, esse método proporciona medidas individuais l_i, de modo a indicar quais observações contribuem mais para l_{max}. Os valores individuais $l_i = \left(1-\frac{\left(\mathbf{e}_i^*\right)^2}{\hat{\sigma}^2}\right)$ referentes ao grupo 2 encontram-se na Tabela 6.4 e na

Figura 6.4. Por meio delas, detectamos quatro observações cujos valores l_i são perceptivelmente maiores que as demais: 21, 38, 44 e 46, sendo que duas delas, a saber, as observações 21 e 44, já haviam sido diagnosticadas pelo método de mínimos quadrados.

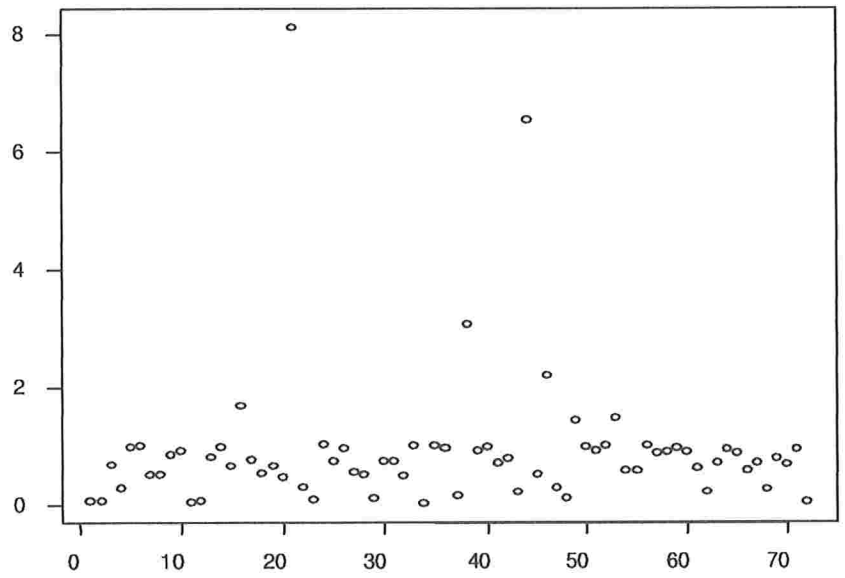


Figura 6.4 - Medidas Individuais l_i - Grupo 2

Tabela 6.4 – Valores absolutos de l_i para os dados do Grupo 2

Caso	$\|l_i\|$	Caso	$\|l_i\|$	Caso	$\|l_i\|$	Caso	$\|l_i\|$	Caso	$\|l_i\|$
1	0,06	16	1,69	31	0,74	**46**	**2,18**	61	0,62
2	0,06	17	0,77	32	0,48	47	0,27	62	0,22
3	0,68	18	0,52	33	1,00	48	0,11	63	0,70
4	0,29	19	0,65	34	0,03	49	1,43	64	0,94
5	0,98	20	0,47	35	1,00	50	0,99	65	0,87
6	1,00	**21**	**8,10**	36	0,96	51	0,91	66	0,57
7	0,52	22	0,29	37	0,14	52	1,01	67	0,71
8	0,52	23	0,09	**38**	**3,07**	53	1,46	68	0,25
9	0,86	24	1,03	39	0,91	54	0,57	69	0,79
10	0,92	25	0,73	40	0,99	55	0,57	70	0,68
11	0,04	26	0,96	41	0,70	56	0,99	71	0,93
12	0,07	27	0,56	42	0,80	57	0,87	72	0,04
13	0,82	28	0,51	43	0,21	58	0,90		
14	0,98	29	0,11	**44**	**6,53**	59	0,96		
15	0,67	30	0,75	45	0,51	60	0,90		

- **Grupo 3:**

Com relação ao grupo 3, as estimativas dos parâmetros do modelo obtidas via mínimos quadrados estão organizados na Tabela 6.5.

Tabela 6.5 - Ajuste do Modelo por Mínimos Quadrados - Grupo 3

	Estimativas	Erro Padrão	Valor da estatística t	p-valor
(Intercepto)	3,27	0,07	47,28	<2e-16
B4	-0,07	0,03	-2,47	0,0161
SIGMA	0,60	0,55	1,10	0,2754
F	-0,98	0,57	-1,73	0,0890
R	-1,45	0,60	-2,40	0,0192
LOG.PAPP	0,30	0,04	6,82	3,41e-09

Nesse ajuste é possível perceber que os coeficientes das variáveis SIGMA e F não são estatisticamente significantes ao nível de 0,05. Os FIV's associados a cada variável foram obtidos com a finalidade de verificar a existência de multicolinearidade e encontram-se na Tabela 6.6.

Tabela 6.6 - Fator de Inflação da Variância - Procedimento de Regressão de MQ - Grupo 3

Variável	B4	SIGMA	F	R	LOG.PAPP
FIV	3,92	74,15	22,68	54,62	5,68

Desse modo, podemos perceber que as variáveis SIGMA, F, R e LOG.PAPP são altamente correlacionadas com uma ou mais variáveis independentes já que apresentam FIV maior que 5.

Os autovalores da matriz $(\mathbf{X}'\mathbf{X})$ são dados por: 3,1756; 1,0058; 0,6851; 0,1267 e 0,0066. Logo, $\kappa = \frac{\lambda_{max}}{\lambda_{min}} = \frac{3,1756}{0,0066} = 481,15$. Como $100 < \kappa < 1000$, podemos concluir pela existência de forte multicolinearidade nos dados.

Como forma de contornar o problema da multicolinearidade, um modelo de regressão em cristas será ajustado. Mas, para isso, o traço será analisado como critério de escolha para o valor para k, tomando k variando de zero a dois.

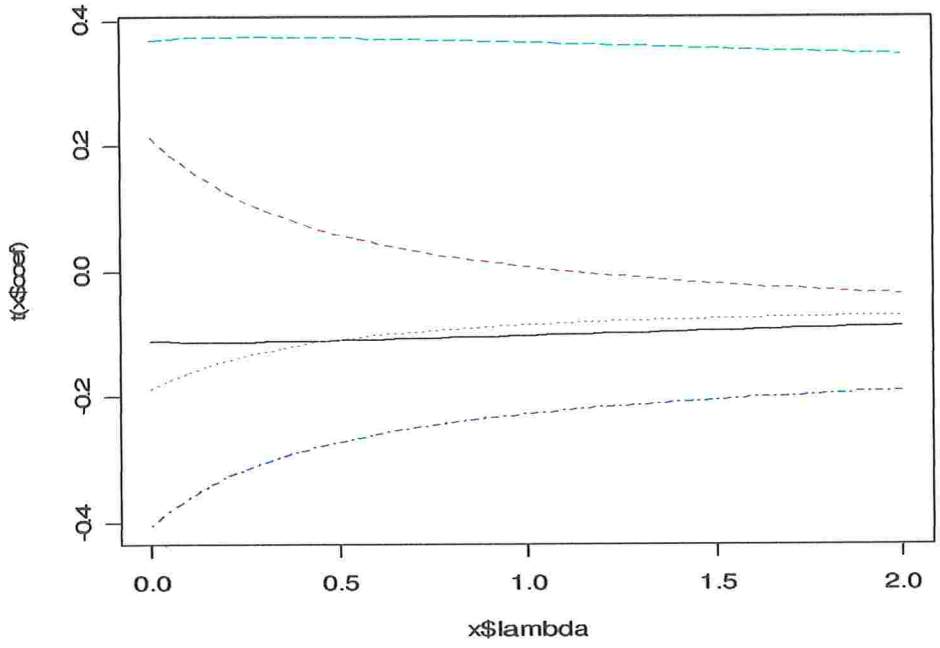


Figura 6.5 – Traço para as variáveis: B4, SIGMA. F, R e LOG.PAPP - Grupo 3

Através da Figura 6.5, podemos perceber que a partir de $k = 1$, os coeficientes tendem a se estabilizar. Assim, esse valor será escolhido e um modelo de regressão em cristas para os dados do grupo 3 pode ser expresso por.

$$\hat{Y} = 3,27 - 0,07 \cdot \text{B4} + 0,02 \cdot \text{SIGMA} - 0,44 \cdot \text{F} - 0,81 \cdot \text{R} + 0,30 \cdot \text{LOG.PAPP}$$

Os valores dos FIV's para as variáveis independentes do grupo 3, após ajuste de um modelo de regressão em cristas, decresceram consideravelmente e são dados na Tabela 6.7.

Tabela 6.7 - Fator de Inflação da Variância - Procedimento de Regressão em Cristas - Grupo 3

Variável	B4	SIGMA	F	R	LOG.PAPP
FIV	0,04	0,11	0,05	0,09	0,06

Calculando-se os elementos da diagonal principal de $\mathbf{H}^* = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^*\right)^{-1}\mathbf{Z}'$ temos, também, que as observações 64, 65 e 66 foram consideradas como as mais influentes, com valor $h_i^* = 0{,}172$. Esse fato pode ser verificado por meio da Figura 6.6. Ao investigarmos suas características podemos perceber que esses três ratos, assim como no grupo 2, apresentaram os maiores valores de SIGMA (0,72) e foram os únicos a apresentarem valores negativos em LOG.PAPP (-0,70). Mas, além disso, apresentaram os maiores valores da variável F (0,54) e da variável R (0,22).

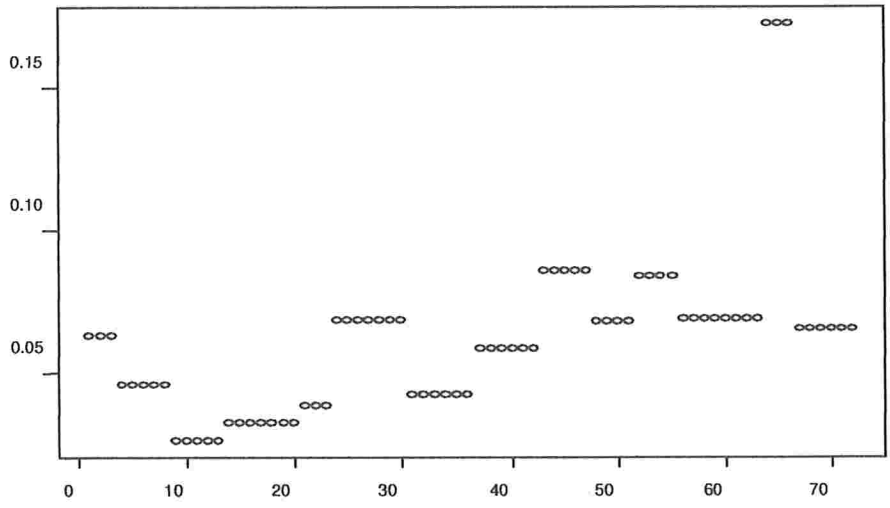


Figura 6.6 - Valores da diagonal principal da matriz H* - Grupo 3

Considerando o modelo de regressão em cristas ajustado para os dados do grupo 3, também não foram detectados pontos influentes por meio da medida D_i^*. Mas se o procedimento adotado fosse pelo método de mínimos quadrados, sete seriam as observações influentes: 21, 44, 64, 65, 69, 70 e 71.

A curvatura máxima obtida foi $C_{\max} = \frac{2 \cdot \lambda_{\max}^*}{\hat{\sigma}^2} = \frac{2 \cdot 0{,}0525435}{0{,}03663602} = 2{,}87$. Dessa forma, alguma sensibilidade local existe nos dados, de acordo com o critério de Cook $(C_{\max} > 2)$. Mas, pelo critério sugerido por Loynes (1997), $C_{\max} = 2{,}87 < 10$, pelo fato de $p = 5$. Desse modo, a partir dos critérios de Cook e Loynes, podemos concluir por uma sensibilidade moderada.

O autovetor associado a $\lambda_{\max}^*$ referente aos dados do grupo 3 sugere como influentes os ratos de números: 44, 43, 46, 49, 21, 48 e 45, todos com componente de $\lambda_{\max}^*$ e maiores que $|0{,}2|$, como pode ser verificado na Figura 6.7.

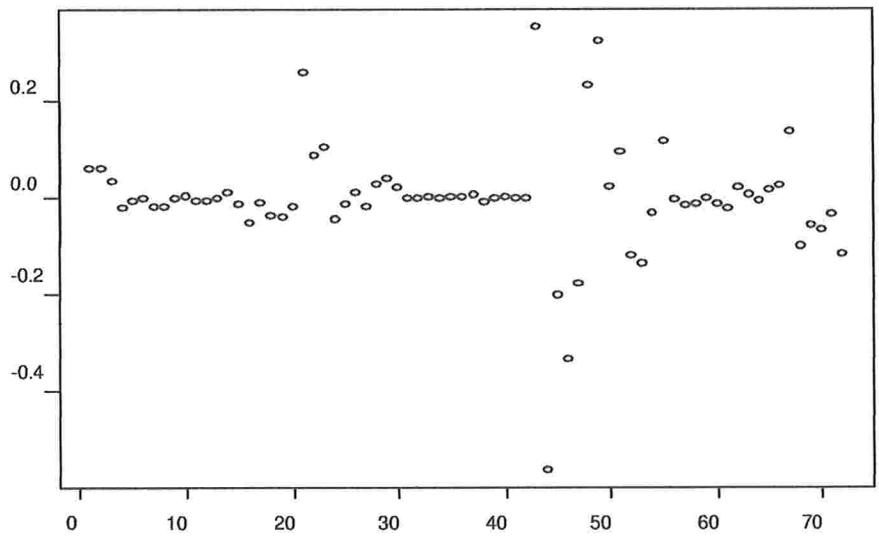


Figura 6.7 - Análise da Influência pelas componentes do autovetor associado a $\lambda_{\max}^*$ - Grupo 3

A inclinação máxima l_{max} foi estimada, para o grupo 3, obtendo-se

$$\left|\nabla LD^*_{\Re}\right| = \left[\sum_{i=1}^{n}\left(1-\frac{(\mathbf{e}_i^*)^2}{\hat{\sigma}^2}\right)^2\right]^{1/2},$$

$$\left|\nabla LD^*_{\Re}\right| = 13{,}53.$$

Assim, como $l_{max} = 13{,}53 < 16{,}46$, temos que l_{max} não sugere sensibilidade local para os dados do grupo 3. Os valores individuais l_i referentes ao grupo 3 encontram-se na Tabela 6.2 e na Figura 6.8. Com base nelas, detectamos quatro observações que são perceptivelmente maiores que o restante dos dados: 21, 38, 44 e 49, sendo que as observações 21, 38 e 44 já haviam sido detectadas no Grupo 2. Acrescenta-se, ainda, que, as observações 21 e 44 também foram identificadas pelo método de mínimos quadrados e pelas componentes de λ^*_{max}.

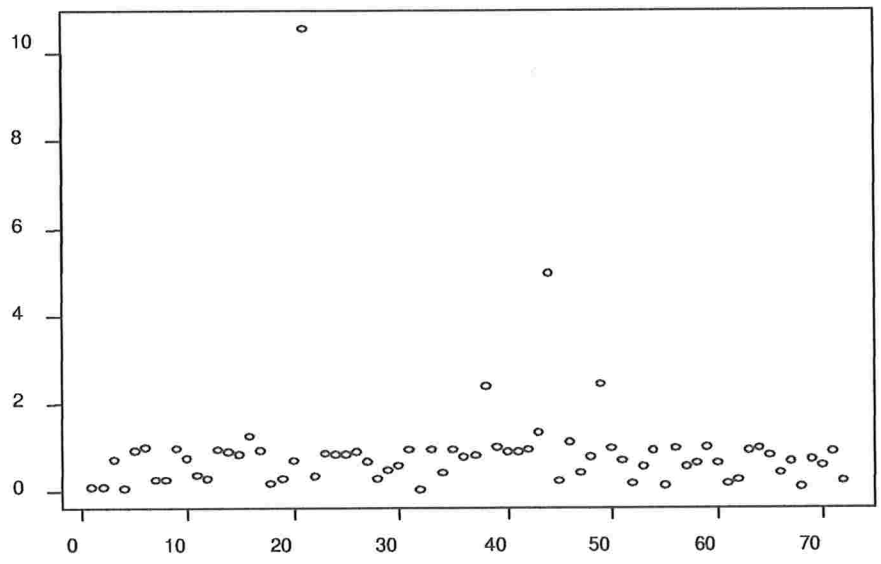


Figura 6.8 - Medidas Individuais l_i - Grupo 3

Tabela 6.8 – Valores absolutos individuais l_i para os dados do Grupo 3

Caso	$\lvert l_i \rvert$	Caso	$\lvert l_i \rvert$	Caso	$\lvert l_i \rvert$	Caso	$\lvert l_i \rvert$	Caso	$\lvert l_i \rvert$
1	0,10	16	1,25	31	0,94	46	1,10	61	0,16
2	0,10	17	0,91	32	0,04	47	0,41	62	0,27
3	0,72	18	0,17	33	0,95	48	0,77	63	0,92
4	0,05	19	0,29	34	0,41	**49**	**2,43**	64	0,97
5	0,91	20	0,69	35	0,95	50	0,98	65	0,80
6	0,99	**21**	**10,5**	36	0,76	51	0,70	66	0,42
7	0,24	22	0,32	37	0,81	52	0,18	67	0,68
8	0,24	23	0,85	**38**	**2,39**	53	0,54	68	0,10
9	0,97	24	0,82	39	1,00	54	0,91	69	0,72
10	0,74	25	0,83	40	0,88	55	0,12	70	0,59
11	0,37	26	0,89	41	0,89	56	0,96	71	0,90
12	0,28	27	0,68	42	0,95	57	0,57	72	0,23
13	0,95	28	0,29	43	1,34	58	0,63		
14	0,89	29	0,46	**44**	**4,96**	59	0,99		
15	0,84	30	0,60	45	0,24	60	0,63		

Com base nas medidas individuais l_i, foram identificadas as observações 21, 38, 44 e 46 no Grupo 2, ou seja, são as que mais contribuem para a inclinação máxima $l_{\max}$. Já no Grupo 3, essas observações são as de número 21, 38, 44 e 49.

Eliminadas essas observações, os modelos selecionados para os grupos 2 e 3, agora reduzidos, encontram-se na Tabela 6.9.

Tabela 6.9 - Modelos de Regressão em Cristas com Observações Eliminadas nos Grupos 2 e 3

Grupo 2 - reduzido	$\hat{Y} = 3{,}17 + 0{,}06B4 - 0{,}62SIGMA + 0{,}25LOG.PAPP - 0{,}15MR4$
Grupo 3 - reduzido	$\hat{Y} = 3{,}26 - 0{,}05B4 - 0{,}04SIGMA - 0{,}37F - 0{,}72R + 0{,}27LOG.PAPP$

Após o ajuste desses dois modelos de regressão em cristas verificamos, ainda, que, para os dois grupos reduzidos considerados, os valores do FIV (que já eram baixos) decresceram um pouco mais. Isso significa que a multicolinearidade foi sensivelmente reduzida precisamente em função da eliminação das observações que mais contribuíam para a inclinação máxima $l_{\max}$, fato que pode ser constatado na Tabela 6.10.

Tabela 6.10 - Comparação de Valores FIV

FIV do Modelo de Regressão em Cristas do Grupo 3 (n=72)				
B4	SIGMA	F	R	LOG.PAPP
0,04	0,11	0,05	0,09	0,06
FIV do Modelo de Regressão em Cristas do Grupo 3 Reduzido (n=68)				
B4	SIGMA	F	R	LOG.PAPP
0,04	0,06	0,03	0,05	0,05
FIV do Modelo de Regressão em Cristas do Grupo 2 (n=72)				
B4	SIGMA	LOG.PAPP	MR4	
0,09	0,02	0,04	0,09	
FIV do Modelo de Regressão em Cristas do Grupo 2 Reduzido (n=68)				
B4	SIGMA	LOG.PAPP	MR4	
0,05	0,02	0,04	0,05	

Os valores para a inclinação máxima $l_{\max}$ foram novamente obtidos para os grupos reduzidos 2 e 3 e seus valores são, respectivamente, 8,2 e 7,7. Pelo fato de serem ambos menores do que a raiz quadrada de $2n + 4\sqrt{14n}$, que nesse caso é 16,1, continuam não sugerindo sensibilidade local nos dados.

Podemos concluir, com base no que foi exposto no presente capítulo, que as técnicas de regressão em cristas foram eficientes no que diz respeito ao seu objetivo principal, qual seja, o de contornar o problema da multicolinearidade existente no conjunto de dados.

É possível concluir, ainda, que ao utilizar medidas de influência adaptadas para modelos de regressão em cristas, a quantidade de observações influentes, obtida pelos diversos métodos empregados, geralmente diminuiu quando comparada com a quantidade de observações influentes obtidas por mínimos quadrados.

Capítulo 7

Considerações Finais

Com base na abordagem descrita nesse trabalho, para o ajuste de um modelo de regressão na existência de multicolinearidade, sugerimos a realização dos seguintes passos:

1) Verificar se a multicolinearidade é gerada por pontos discrepantes. Em caso positivo, o uso de estimadores viciados, em particular o estimador em cristas, pode não ser uma alternativa eficaz. Para saber se a multicolinearidade é gerada por pontos discrepantes, gráficos de dispersão dos componentes principais normalizados correspondentes aos maiores autovalores da matriz de correlação devem ser elaborados. Esses gráficos devem apresentar um comportamento aleatório dos pontos caso os dados não contenham multicolinearidade ou pontos extremos nas variáveis explicativas.

Caso se conclua que a multicolinearidade é devida à natureza das variáveis e se decida utilizar o procedimento de regressão em cristas para solucionar o problema, sugere-se:

2) Ajustar o modelo de regressão em cristas, calcular as medidas D_i^* e ou $DFFITS^*(i)$ para cada observação e identificar aquelas que são influentes;

3) Com o ajuste realizado é possível efetuar a análise das medidas de influência local $C_{\max}$ e $l_{\max}$. Além disso, podemos, também, medir a contribuição individual de cada l_i na inclinação máxima $l_{\max}$.

Detectados pontos influentes, caso se decidida pela não eliminação das observações, sugerimos o uso das técnicas de regressão robusta em cristas, apresentadas no Capítulo 5.

A elaboração de um programa para o ajuste da técnica de regressão robusta em cristas fica como sugestão para um trabalho futuro.

Finalizando, gostaríamos ainda de destacar o recente artigo de Labra, Aoki e Rojas (2007). Os autores propõem medidas de influência local para o estimador em cristas em modelos cujos erros têm distribuição elíptica. Tal análise, fora do contexto do nosso trabalho que é restrito a modelos com erros com distribuição Normal, seria um interessante tópico de pesquisa futura.

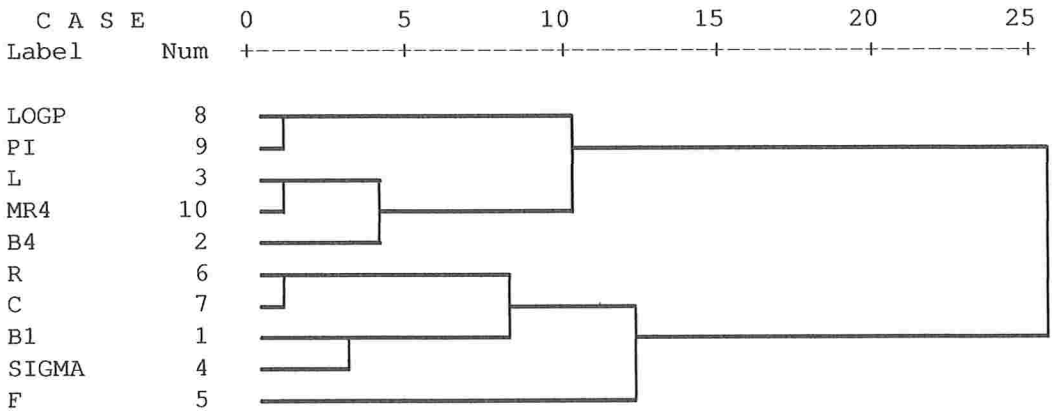
Apêndice A

Tabela A.1 - Matriz de correlações

	POTENCIA	B1	B4	L	SIGMA	F	R	C	LOGP	PI	MR4
POTENCIA	1,00000										
B1	-,52297	1,00000									
B4	,62254	-,26457	1,00000								
L	,59009	-,32040	,96832	1,00000							
SIGMA	-,81348	,80866	-,53280	-,49920	1,00000						
F	-,50114	,39930	-,15120	-,04275	,54931	1,00000					
R	-,71070	,70208	-,55751	-,56589	,86196	,06856	1,00000				
C	-,71783	,77981	-,62815	-,60560	,92954	,28694	,94048	1,00000			
LOGP	,86043	-,41238	,80174	,79031	-,66814	-,48965	-,54307	-,59914	1,00000		
PI	,82978	-,35617	,77351	,76025	-,61331	-,53433	-,44876	-,51699	,99098	1,00000	
MR4	,57726	-,28404	,98041	,98715	-,50182	-,11206	-,52023	-,58559	,77908	,75597	1,00000

Tabela A.2 - Dendograma da Análise de Agrupamento

Método: Centróide
Medida de similaridade: Correlação linear de Pearson



Apêndice B

Apresentaremos a seguir os comandos utilizados no software R. Os pacotes necessários foram: *car* e *MASS*. A título de ilustração, todas as rotinas a serem apresentadas aqui correspondem àquelas utilizadas para as análises do grupo 2. A diferença para o grupo 3 reside apenas na troca de variáveis.

- Calcular os valores do FIV associado a cada variável a de um modelo:

```
> vif(lm(POTENCIA~B4+SIGMA+LOG.PAPP+MR4, dados=ratos))
```

- Padronização de matrizes:

```
> x<-scale(X, center=TRUE, scale=TRUE)
```

- Matriz inversa de X:

```
> solve(X)
```

- Obter resíduos de um modelo de regressão de mínimos quadrados:

```
> residuals(lm(POTENCIA~B4+SIGMA+LOG.PAPP+MR4, dados=ratos))
```

- Contrução do Traço para a escolha de k, com k variando de 0 a 2:

```
> plot(lm.ridge(POTENCIA~B4+SIGMA+LOG.PAPP+MR4, ratos, lambda=seq(0, 2,
0.05)))
```

- Determinação do modelo de regressão em cristas após a escolha de k :

> lm.ridge(POTENCIA~B4+SIGMA+LOG.PAPP+MR4, ratos, lambda=1.5)

- Obtenção dos resíduos do modelo de regressão em cristas:

1) chamenos de br o vetor dos coeficientes do modelo em cristas estimados;

> br<-matrix(c(b0, b1, b2, b3, b4), nrow=5)

2) Criar um vetor de U de tamanho 72.

> U<-rep(1,72)

3) Dispor os vetores de interesse em uma matriz X.

> X<-cbind(U, B4, SIGMA, LOG.PAPP, MR4)

3) Escrever a função dos resíduos em cristas.

> eridge<-POTENCIA-X%*%br

- Obtenção da inclinação máxima:

> eridge2<-eridge^2

> grad<-(U-(eridge2/s2))

> grad2<-grad^2

> grad3<-sum(grad2)

> grad4<-sqrt(grad3)

- Medidas individuais l_i da inclinação máxima:

> l1<-(1-(eridge2[1,]/s2))

- Estimador de mínimos quadrados de σ^2 :

> res<-c(residuals(modelo))

> res2<-res^2

> s2<-sum(res2)/(72-5)

- Curvatura máxima:

> D<-diag(c(eridge))

> m<-(D%*%X%*%solve(t(X)%*%X+1.5*diag(1,5))%*%t(X)%*%D)

> eigen(m)$values

OBS: O maior autovalor foi obtido colocando em ordem crescente os autovalores por meio do comando:

> sort(eigen(m)$values)

- Matriz "hat" na regressão em cristas:

> H<-X%*%solve((t(X)%*%X+1.5*diag(1,5)))%*%t(X)

> h<-diag(H)

- FIV na regressão em cristas:

> vifr<-solve(t(X)%*%X+1.5*diag(1,4))%*%t(X)%*%X%*%solve(t(X)%*%X+1.5*diag(1,4))

> vifridge<-diag(vifr)

- Cálculo de D_i^*:

> Di<-(1/(5*1.5))%*%t(b-bi)%*%t(X)%*%X%*%(b-bi),

sendo que b representa o vetor dos estimadores em cristas com todas as observações e bi representa o vetor dos estimadores em cristas sem a *i-ésima* observação.

- Obtenção do vetor dos estimadores em cristas bi sem a *i-ésima* observação:

> lm.ridge(POTENCIA~B4+SIGMA+LOG.PAPP+MR4,ratos, subset=c(1:i-1,i+1:72), lambda=1.5)

Bibliografia

Andrade, F. C. (2004). "Pontos de Alavanca em Regressão". *Dissertação de Mestrado*. IME-USP.

André, C. D. S. de, Elian, S. N., Bruscato, A. (1997). "Relatório de Análise Estatística Sobre o Projeto: Relação Estrutura-Atividade de Anestésicos Locais N,N [Dimetilamina] Etil Benzoatos Para-Substituídos". RAE-CEA-9710.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). "Regression Diagnostics". New York: John Wiley.

Billor, N. and Loynes, R. M. (1993). "Local Influence: A New Approach", *Communications in Statistics – Theory and Methods*, 22, 1595-1611.

Billor, N. and Loynes, R. M. (1999). "An Application Local Influence Approach to Ridge Regression". *Journal of Applied Statistics*. Vol. 26, nº 2, 1999, 177-183.

Cook, R. D. (1977). "Detection of Influential Observations in Linear Regression". *Technometrics*, 19, 15-18.

Cook, R. D. (1986). "Assessment of Local Influence (wit discussion)". *Journal of the Royal Statistical Society*, Series B, 48, 133-169.

Dorsett, D. e Gunst, R. F. (1982). "Bounded-Leverage Weights for Robust Regression Estimators". Technical Report 171, Southern Methodist University, Dept. of Statistics.

Labra, F. V., Aoki, R., Rojas, F. (2007). "An Application of the Local Influence Diagnostics to the Ridge Regression Under Elliptical Model". *Communications in Statistics – Theory and Methods*, 36, 767-779.

Hill, R. W. (1977). "Robust Regression When There Are Outliers in the Carriers". Unpublished Ph.D. dissertation, Harvard University, Dept. of Statistics.

Hocking, R. R. (1984). "Discussion". *Technometrics*, 26, 321-323.

Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). "Ridge Regression: Some Simulations". *Communications in Statistics*, 4, 105-123.

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems". *Technometrics*, 12, 55-67.

Hogg, R. V. (1979). "An Introduction to Robust Estimation". *Robustness in Statistics (R. L. Launer & G. N. Wilkison, eds.)*. New York: Academic Press, 1-17.

Labra, F. V., Aoki, R. e Rojas, F. (2007). "An Application of the Local Influence Diagnostics to Ridge Regression Under Elliptical Model". *Communications in Statistics – Theory and Methods* 36, 767-779.

Lawless, J. and Wang, P. (1976). "A Simulation Study of Ridge and Other Regression Estimators". *Communications in Statistics – Theory and Methods*. A 5, 307-323.

Lawrence, K. D. and Marsh, L. C. (1984). "Robust Ridge Estimator Methods for Predicting U.S. Coal Mining Fatalities" *Communications in Statistics*. A.13, 139-149.

Lichtenstein, C. and Velleman, P. F. (1983). "The Effects of Ridge Regression on High Leverage Points in the Data", unpublished manuscript.

Lindley, D. V. and Smith, A. F. M. (1972). "Bayes Estimate for the Linear Model". *Journal of the Royal Statistical Society*, Series B 34, 1-41.

Longley, J. W. (1967). "An Appraisal of Least Squares Programs for the Electronic Computer From the Point of View of the User". *Journal of the American Statistical Association*, 62, 819-841.

Loynes, R. M. (1997). "A New Measure in Local Influence", *Research Report 474/97*, School of Mathematics and Statistics, The University of Sheffield.

Marquardt, D. W. (1970). "Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation". *Technometrics*, 12, 591-612.

Mason, R. L. and Gunst, R. F. (1985). "Outlier-Induced Colinearities". *Technometrics*, 27, 401-407.

Montgomery, D. C. and Peck, E. A. (1982). "Introduction to Linear Regression Analysis", New York: John Wiley.

Morisson, D. F. (1976). "Multivariate Statistical Methods". Second Edition. Editora McGraw-Hill Kogakusha, LTD.

Oishi, J. (1983). "Regressão Sobre Cristas". *Dissertação de Mestrado*. IME-USP.

Ramsay, J. O. (1977). "A Comparative Study of Several Robust Estimates of Slope, Intercept and Scale in Linear Regression." *J. Amer. Statist. Assoc.*, 72, 608-618.

Tripp, R. E. (1983). "Nonstochastic Ridge Regression and Effective Rank of the Regressors Matrix". Unpublished Ph.D. dissertation, Virginia Polytechnic Institute and State University, Dept. of Statistics.

Vinod, H. D. and Ullah, A. (1981). "Recent Advances in Regression Methods". New York: Marcel Dekker, inc.

Walker, E. and Birch, J. B. (1988). "Influence Measures in Ridge Regression". *Technometrics*, 30, 221-227.

Weisberg, S. (1980). "Applied Linear Regression". *Wiley Series in Probability and Statistics*. New York.